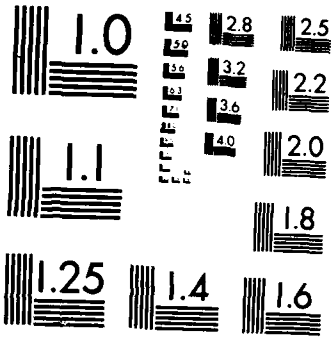
AD-A175 028 ASYMPTOTIC PROPERTIES OF DISTRIBUTED AND COMMUNICATING 1/1
STOCHASTIC APPROXI (U) BROWN UNIV PROVIDENCE RI
LEFSCHETZ CENTER FOR DYNAMICAL SYSTE
UNCLASSIFIED H J KUSHNER ET AL FEB 86 LCDS-86-11 F/G 12/1 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Asymptotic Properties of Distributed and Communicating Stochastic Approximation Algorithms

by

Harold J. Kushner and G. Yin

February 1986 LCDS #86-11

Lefschetz Center for Dynamical Systems

...d Applied Mathematics Brown University Providence RI 02912

Asymptotic Properties of Distributed and Communicating Stochastic Approximation Algorithms

by

Harold J. Kushner and G. Yin

February 1986 LCDS #86-11

86-12-11-121

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 86-2140	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Asymptotic Properties of Distributed and Communicating Stochastic Approximation Algorithms		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) H.J. Kushner and G. Yin		8. CONTRACT OR GRANT NUMBER(s) AFOSR-81-0116
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lefschetz Center for Dynamical Systems Division of Applied Mathematics Brown University, Providence, RI 02912		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304 A4
11. CONTROLLING OFFICE NAME AND ADDRESS AFOSR Bolling Air Force Base Washington, DC 20332 nm		12. REPORT DATE Feb 1986 13. NUMBER OF PAGES 54
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) same as 11		15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release: distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Approved for public release;
distribution unlimited.

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Included

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Asymptotic Properties of Distributed and Communicating Stochastic Approximation Algorithms

by

Harold J. Kushner*
G. Yin+

Lefschetz Center for Dynamical Systems
Division of Applied Mathematics
Brown University
Providence, RI 02912

February 1986

* This research was supported in part by the Air Force Office of Scientific Research under Grant #AFOSR-81-0116; and Office of Naval Research Grant #N00014-83-K-0542.

+ This research was supported in part by the Army Research Office under Grant #DAAG-29-84-K-0082; and the Office of Naval Research under Grant #N00014-85-K-0607.

Abstract

The asymptotic properties of extensions of the type of distributed or decentralized stochastic approximation proposed ~~in [1]~~ are developed. Such algorithms have numerous potential applications in decentralized estimation, detection and adaptive control, or in decentralized Monte Carlo simulation for system optimization (where they can exploit the possibilities of parallel processing). The structure involves several isolated processors (recursive algorithms) who communicate to each other asynchronously and at random intervals. The asymptotic (small gain) properties are derived. The communication intervals need not be strictly bounded and they and the system noise can depend on the (communicating) system state. State space constraints are also handled. In many applications, the dynamical terms are merely indicator functions, or have other types of discontinuities. The 'typical' such case is also treated, as is the case where there is noise in the communication. The linear stochastic differential equation satisfied by the (interpolated) asymptotic normalized error sequence is derived, and issued to compare alternative algorithms and communication strategies. Weak convergence methods provide the basic tools.

1. Introduction

In [1], [2], Tsitsiklis proposed a very interesting model for a decentralized (distributed) recursive algorithm of the stochastic approximation (SA) type, with only asynchronous communications between the separate processors, and developed a scheme for proving w.p.1 convergence. That work appears to be the first of its type - for the decentralized SA problem. Such distributed algorithms are of rapidly growing interest. Various potential applications in adaptive control, estimation and in communcation networks were proposed; e.g., several processors might do an identification of the parameters of an identical linear system (but with different inputs) and occasionally (asynchronously) share their latest estimates, or several processors might do monte carlo simulations of the SA type to locate the minimum of a regression function, and occasionaly share their estimates. There are two main purposes for algorithms of the type discussed here and in [1]: to exploit the opportunities provided by parallel processing for monte-carlo methods of system optimization or evaluation; situations in which there are physically separate systems (estimators, trackers, controllers) which act on or follow essentially the same physical system - and which occasionally communicate to take advantage of the 'others' information.

The assumptions in [1] were fairly strong with respect to the great variety of potential applications, and the method of analysis required numerous detailed estimates. We analyze essentially the same algorithm here. In addition to getting the basic convergence results, our methods can handle the constrained (projected) algorithm, the case where the communication intervals and noise depend on the state, the general rate of convergence problem, and the case where there is communication noise. Instead of letting the 'gain' parameter go to zero as $n \to \infty$ (as frequently done in classical SA) we keep it a constant, and work with convergence in the sense of weak convergence. There are several reasons for this. First, when working with practical systems the chosen gains almost never go to zero, since one usually wants an algorithm that can track slow changes and is robust with respect to large bursts of noise. Our method can be adapted to get weak and even w.p.1 convergence when the gains do go to zero, and we comment on this in Section 8. Even if the gains do go to zero, w.p.1 convergence is not much more useful or interesting than weak convergence. Weak convergence methods locate the points where the process spends most time (asymptotically), and as time goes to ∞, an increasing (to one) proportion of time is spent arbitrarily close to such points. Then, one can often use the powerful 'large deviations' methods to show, under very broad conditions, that ultimate escape from a small neighborhood of such points is impossible (when the gains go to zero) [3], [4]. Alternatively, once the weak convergence methods have located the 'stable points', perturbed Liapunov methods such as that in [10] can often be used to get w.p.1 convergence. One of the key questions in the analysis of any algorithm is the rate of convergence (the asymptotic normalized variance), and the analysis of the 'rate' is almost always done via weak convergence methods. General background and applications in many areas are in [6] to [8]. Weak convergence methods are also <u>much</u> <u>easier</u> to use than the standard w.p.1 oriented methods; in many cases, a valid result can be obtained almost by inspection. This and the wide variety of problems which can be handled make it a more widely useful tool than 'w.p.1' methods. The symbol $\Rightarrow$ is used to denote weak convergence, and some definitions and properties of this convergence are stated in the Appendix 1.

The methods used here are quite efficient. Problems with potentially unbounded intercommunication intervals (e.g., where the interval is geometrically distributed) can be handled. We can also treat important cases where the dynamics are discontinuous or where the communication intervals and system noise depend on the system state, or where there are state space constraints. The case of discontinuous dynamics is of considerable importance in applications: often an estimate increases or decreases by a fixed amount ϵ - depending simply on whether a certain event occurred or not. Similarly, for state dependent communication times; a processor might want to communicate if either a given amount of time has passed since the last communication or if the state of the processor has changed by more than a given amount. In many applications (e.g., the decentralized form of the automata routing problem in [5]) the noise is naturally state dependent.

A theory of 'rate of convergence' is also developed, which allows an objective comparison among alternative algorithms. Using this, in Sections 6 and 7, we comment on and compare the behavior of the algorithm with the centralized and various 'deterministically' decentralized forms, in order to get a better understanding of its behavior, and to see what are the preferable communication strategies. We can also allow ‘noise’ in the communication, such as might be the case if the processors were physically separated and communicated via a noisy radio link. See Section 7.

The basic algorithm will be described next. Section 2 contains a ‘technical’ estimate which will be useful in the sequel. Section 3 deals with the basic weak convergence result in the function spaces $D[0,\infty)$ or $C[0,\infty)$ (see Appendix for the definitions), and shows that a suitable continuous time interpolation $X^{\epsilon}(\cdot)$ of the iterates $\{X_n\}$ converges weakly to the solution of a certain ODE as the gain parameter $\epsilon \to 0$. The state dependent noise/intercommunication time case and the discontinuous dynamics case are also treated there. Section 4 concerns a ‘projection’ algorithm to handle state space constraints. Here, the limit satisfies a ‘projected’ ODE. The asymptotics of $X^{\epsilon}(t_{\epsilon}+\cdot)$ are dealt with in Section 5, where $t_{\epsilon} \to \infty$ as $\epsilon \to 0$. This yields the ultimately desired result concerning the location of the iterates for large n and small ϵ. Finally, the rate of convergence and comparison with a centralized processor is developed in Sec.ion 6 and 7. A discussion of some of the probable advantages and uses of the algorithm appears in Section 7. Section 8 contains a comment on the case where ϵ is replaced by $\epsilon_n \to 0$.

<u>The basic algorithm</u>. We assume that there are q parallel processors, each with a state variable of dimension r. Let X_n^i denote the state of processor i at time n and define $X_n = (X_n^1, ..., X_n^r)$. The symbol X generally denotes a qr-vector which we partition as $X = (X^1, ..., X^q)$, where each X^i is

an r-vector. The 'observation' of processor i at time n is $b^i(X^i_n,\xi^i_n)$, where ξ^i_n is the 'noise'. Write $\xi_n = (\xi^1_n, ..., \xi^q_n)$, $\xi = (\xi^1, ..., \xi^q)$, and $B(X,\xi) = (b^1(X^1,\xi^1), ..., b^q(X^q,\xi^q))$. Write $b^i(X^i,\xi^i) = (b^i_1(X^i,\xi^i), ..., b^i_r(X^i,\xi^i))$, the $b^i_k(\cdot)$ being scalar valued. (All the above vectors are column vectors.) For vectors X^i in E^r, we often write simply x.

Let $\{A_n\}$ be a sequence of (possibly random) qr × qr matrices, where A_n can be written in the form

$$A_n = \begin{bmatrix} a_{11}(n) & \cdots & a_{q1}(n) \\ & & \\ a_{1q}(n) & \cdots & a_{qq}(n) \end{bmatrix},$$

where each $a_{ij}(n)$ is a <u>diagonal</u> r × r matrix with non-negative entries and $\Sigma_i a_{ij}(n) = I_r$ the identity matrix in E^r, Euclidean r-space (i.e., the 'matrix valued' rows of A_n are 'convexifying'). Suppose that there is a scalar $\alpha_0 > 0$ such that $a_{ii} \geqslant \alpha_0 I_r$ and, for $i \neq j$, either $a_{ij}(n) = 0$ or else $a_{ij}(n) \geqslant \alpha_0 I_r$.

The algorithm is

$$\begin{aligned} X^i_{n+1} &= \sum_j a_{ji}(n)X^j_n + \epsilon b^i(X^i_n,\xi^i_n) \\ X_{n+1} &= A_n X_n + \epsilon B(X_n,\xi_n). \end{aligned} \tag{1.1}$$

At time n, processor i (i = 1,...,q) decides whether or not to communicate the current value of its state to any other processor and takes an observation $b^i(X^i_n,\xi^i_n)$. If there is no communication <u>to</u> processor i, then we set $a_{ii}(n) = I_r$ and $a_{ji}(n) = 0$ for $j \neq i$, and the iteration (for processor i at time n) is of the standard SA type: $X^i_{n+1} = X^i_n + \epsilon b^i(X^i_n,\xi^i_n)$. If there are any communications to processor i from some processors $j \neq i$ at time n, then for such communicating processors j, $a_{ji}(n) \geqslant \alpha_0 I_r$ and the updated state X^i_{n+1} for processor i is a convex combination of X^i_n and of the states X^j_n communicated to it, added to its own SA increment $\epsilon b^i(X^i_n, \xi^i_n)$. The requirement that either (for $j \neq i$) $a_{ji}(n) \geqslant \alpha_0 I_r$ or $a_{ji}(n) = 0$ simply means that if processor j communicates to processor i at time n, processor i can choose to ignore the communication, but if it incorporates the received X^j_n into its own state, it must do so in a 'non-trivial' way. For notational simplicity, we omit the symbol for the ϵ-dependence of X_n.

In [1], the algorithm was slightly more complex, since the dimensions of the X^i were not necessarily the same and a somewhat more complicated block structure of A_n was used. But, with no additional mathematical work (although with a more complex notation), such extensions can readily be incorporated into our framework. It should be clear from the development, that many related algorithms and conditions can be treated by essentially identical methods.

2. Some Preparatory Estimates

This section is devoted to obtaining the rate of convergence of the product $A_n...A_k$ as $n \to \infty$. We use the assumption

C2.1. *Let* F_n *be an increasing sequence of* σ*-algebras such that* F_n *measures* $\{X_i, i \leq n, \xi_i, A_i, i < n\}$. *There are a scalar* $p_0 > 0$ *and integer* m_0 *such that*

$$(2.1) \qquad P_{F_n}\{\text{processor i communicates to processor j on } [n,n+m_0)\} \geq p_0$$

for all n *and* i,j, *and* $i \neq j$.

Remark. In [1], it was assumed that there is an m_0 such that $p_0 = 1$. (C2.1) covers the case where at each instant each processor flips a coin to decide whether to communicate or not. More generally, there often is a process $\{\breve{A}_n\}$ such that $\{\breve{A}_i, \xi_i, i < n, X_i, i \leq n\}$ is Markov, and A_n is a component of $\breve{A}_n$. With this model, if F_n denotes the minimal σ-algebra which measures $\{\breve{A}_i, \xi_i, i < n, X_i, i \leq n\}$, then (C2.1) covers many interesting cases where the inter-communication intervals are not bounded a priori -- and might be 'state' dependent. The condition seems to be unrestrictive.

For $n \geq k$, define $\Phi(n|k) = A_n...A_k$ and set $\Phi(n|n+1) = I_{qr}$, the identity matrix in E^{qr}.

Lemma 2.1. *Assume* (C2.1) *and the conditions on* $\{A_n\}$ *in Section* 1. *Then* $\Phi_k \equiv \lim_n \Phi(n|k)$ *exists* w.p.1 *and for each* $i \leq r$, *all the rows* $i, i+r, ..., i+qr-r$ *of* Φ_k *are equal*. *Also*

$$(2.1a) \qquad E|\Phi(n|k) - \Phi_k| \to 0 \text{ geometrically as } n - k \to \infty,$$

$$(2.1b) \qquad E_{F_k}|\Phi(n|k) - \Phi_k| \to 0 \text{ geometrically as } n - k \to \infty,$$

uniformly in k and ω (w.p.1). Also $E_{\mathcal{F}_n} \Phi(n|k)$ converges to Φ_k geometrically, uniformly in ω, k, as $n \to \infty$.

Remark. The fact that the limit Φ_k exists is almost obvious if we look at the $\{A_n\}$ as transition matrices for a Markov chain.

Proof. The proofs of (2.1a) and (2.1b) are essentially the same and only (2.1a) will be proved. We will evaluate $E|\Phi(n|k) - \Phi_k|$ by a slight variation of the proof of [1, Lemma 5.2.1]. Owing to the block diagonal structure of the A_n, in calculating the product $\Phi(n|k)$, the r sets of rows $(i, i+r, ..., i+qr-r)$, $i \leq r$, do not interact and we can (and will) let $r = 1$ without loss of generality.

The geometric convergence of $\Phi(n|k)$ to Φ_k was proved in [1, Lemma 5.2.1] when $p_0 = 1$ (see remark below (2.4)). By (C2.1), there are $\alpha_1 > 0$ and an increasing sequence of random times $\{N_i\}$ such that the components of $\Phi(N_{2i+1}|N_{2i})$ are all $\geq \alpha_1$. This and the convergence result for $p_0 = 1$ implies that $\Phi(n|k)$ converges w.p.1 to some matrix Φ_k, as $n \to \infty$. All the rows of such a limit must be equal, and the entries of each row must sum to unity. Let $\phi_k(1), ..., \phi_k(q)$ denote the scalar elements of any row of Φ_k, and let the vectors $v_1, ..., v_q$ span E^q, and define $e = (1, 1, ..., 1)$. Define $c(x) = \Sigma\phi_k(i)x_i$ where $x = (x_1, ..., x_q)$. Both e and x are column vectors. Then $\Phi_k x = c(x)e$. All norms here and elsewhere are in the ℓ_∞ sense.

For a matrix M,

$$|M| = \sup_{|x|=1} |Mx| \leq \Sigma |Mv_i|.$$

Thus, we need only show, for any vector x, that $E|\Phi(n|k)x - c(x)e| \overset{n}{\to} 0$ geometrically. Define $x(n|k) = \Phi(n|k)x$. Let $c(n|k)$ denote the minimum value of the components of $x(n|k)$. We can write $x(n|k) = y(n|k) + c(n|k)e$, where all the components of $y(n|k)$ are non-negative and

$$(2.2) \qquad E|\Phi(n|k)x - c(x)e| \leqslant E|y(n|k)| + E|c(x) - c(n|k)|.$$

By the 'convexification' properties of the A_n,

$$(2.3a) \qquad |y(n + 1|k)| \leqslant |y(n|k)|$$

$$(2.3b) \qquad c(n|k) \leqslant c(n + 1|k) \leqslant c(n|k) + |y(n|k)|.$$

By (C2.1), there is an $\alpha_0 > 0$ such that w.p. p_0 (conditioned on F_n) all the elements of $\Phi(n + m_0|n)$ are $\geqslant \alpha_0$. This, together with the 'convexification' property of the A_n implies that

$$(2.4) \qquad E|y(n + m_0|k)| \leqslant (1 - \alpha_0 p_0)E|y(n|k)|.$$

(If $p_0 = 1$, then drop the E in (2.4), and (2.3), (2.4) yield w.p.1 convergence.) The asserted geometric convergence is a consequence of (2.3), (2.4) and the w.p.1 convergence of $\Phi(n|k)$ (hence of $c(n|k)$ to $c(x)$). The last sentence of the theorem follows by a similar argument. Q.E.D.

Remark on Other Cases. One can readily work with the case where all of the processors do not necessarily communicate with each other. We comment only on one special case. Let processors 1, ..., q_1 communicate to each other but not to the other processors, and let processors q_1+1, ..., q, communicate only to processors 1, ..., q_1 but not to each other. Then $\Phi(n|k)$ converges geometrically to a matrix Φ_k which takes the form

$$\Phi_k = \left[\begin{array}{c|ccc} & M^k_{q_1+1,1} & \cdots & M^k_{q1} \\ & & \vdots & \\ 0 & M^k_{q_1+1,q_1} & \cdots & M^k_{qq_1} \\ & I_r & & 0 \\ & & \ddots & \\ & 0 & & I_r \end{array}\right].$$

The i, i+r, ... rows of the upper right hand block are not necessarily equal.

3. Convergence: The Limit ODE

Nonstate Dependent $\{A_n,\xi_n\}$. We will work with several sets of assumptions. First, the basic convergence theorem will be proved when the sequences $\{A_n\}$ and $\{\xi_n\}$ are non state-dependent and independent of each other, and then the restrictions will be weakened. Let E_k denote expectation, conditioned on $\{X_i,\ i \leq n,\ A_i,\ \xi_i,\ i < n\}$. We will use subsets of the following assumptions. Theorem 3.1 is the basic weak convergence theorem, from which most other results will follow. The conditions do not seem to be restrictive.

(C3.1) $\{A_k\}$ *and* $\{\xi_n\}$ *are independent of each other*.

(C3.2) $B(X,\xi) = B_0(X,\hat{\xi}) + B_1(X)\tilde{\xi}$, *where the* $B_i(\cdot)$ *are continuous*, $(B_0(\cdot,\hat{\xi})$, *uniformly in* $\hat{\xi})$, $\{\hat{\xi}_k\}$ *is a sequence of bounded random variables and* $\{\tilde{\xi}_k\}$ *is a sequence with zero mean and bounded 4th moment*.

(C3.3) *There is a continuous function* $\bar{B}(X) \equiv (\bar{b}^1(X^1), \ldots, \bar{b}^q(X^q))$ *such that*

$$E_k B(X,\hat{\xi}_n) - \bar{B}(X) \to 0, \quad E_k \check{\xi}_n \to 0$$

in probability for each X, *as* $n - k \to \infty$.

(C3.4) *There are a matrix* $\bar{\Phi}$ *and a sequence* $m_\epsilon \to \infty$ *such that* $\epsilon m_\epsilon \equiv \delta_\epsilon \to 0$ *and*

$$E\left|\frac{1}{m_\epsilon}\sum_{n}^{n+m_\epsilon-1} E_n\Phi_k - \bar{\Phi}\right| \overset{\epsilon}{\to} 0, \quad \textit{uniformly in } n.$$

Remark and Definition. Under the conditions of Lemma 2.1, $\bar{\Phi}$ must have the form

$$\bar{\Phi} = \begin{bmatrix} \bar{\phi}_1 , \dots , \bar{\phi}_q \\ \\ \bar{\phi}_1 , \dots , \bar{\phi}_q \end{bmatrix} \equiv \begin{bmatrix} \hat{\Phi} \\ \vdots \\ \hat{\Phi} \end{bmatrix}$$

where the $\bar{\phi}_i$ are diagonal matrices with diagonal denoted by $(\bar{\phi}_{i1}, ..., \bar{\phi}_{ir})$ and $\Sigma_j \bar{\phi}_{ji} = 1$. For any vector X we have the form $\bar{\Phi}X = (y, ..., y)$ and $\Phi_k X = (y_k, ..., y_k)$ for some y and y_k in E^r. Let $\hat{\Phi}$ denote the row of r × r matrices $[\bar{\phi}_1, ..., \bar{\phi}_q]$. Let $\bar{B}(x)$ denote $\bar{B}(x,x,...x)$, and $B(x,\xi)$ denote $B(x,x,...,\xi)$.

C3.5. The ODE (3.1) has a unique solution for each initial condition.

$$\begin{matrix} \dot{x}_1 = \bar{\phi}_{11}\bar{b}_1^1(x) + \dots + \bar{\phi}_{q1}\bar{b}_1^q(x) \\ \vdots \\ \dot{x}_r = \bar{\phi}_{1r}\bar{b}_r^1(x) + \dots + \bar{\phi}_{qr}\bar{b}_r^q(x) \end{matrix} = \hat{\Phi}\,\bar{B}(x). \tag{3.1}$$

C3.3'. There are a continuous $\bar{B}(\cdot)$ and $m_\epsilon \to \infty$ such that $\epsilon m_\epsilon \equiv \delta_\epsilon \to 0$ and

$$\frac{1}{m_\epsilon} \sum_n^{n+m_\epsilon -1} E_n B(X, \xi_k) \overset{\epsilon}{\to} \bar{B}(X)$$

in probability for each X, uniformly in n.

C3.4'. There is a matrix $\bar{\Phi}$ such that, as $n - k \to \infty$,

$$E|E_k\Phi_n - \bar{\Phi}| \to 0.$$

Let n_ϵ be a sequence tending to ∞ and such that $\sqrt{\epsilon}n_\epsilon \to 0$, and, for $n \geq n_\epsilon$,

$$\sup_k P\{|\Phi(k + n_\epsilon|k) - \Phi_k| \geq \epsilon^2\} \leq \epsilon^2.$$

There is such a sequence, by Lemma 2.1. In fact, we can use $n_\epsilon = 0(\log 1/\epsilon)$. Define

$$X_0^\epsilon = \Phi(n_\epsilon|0)X_0 + \epsilon \sum_0^{n_\epsilon-1} \Phi_{k+1}B(X_k,\xi_k)$$

and for $t \geq 0$ define $X^\epsilon(\cdot)$ by $X^\epsilon(t) = X_n$ for $t \in [(n-n_\epsilon)\epsilon, (n-n_\epsilon+1)\epsilon)$. Write $X^\epsilon(\cdot) = (X^{\epsilon,1}(\cdot), \ldots, X^{\epsilon,q}(\cdot))$. It will turn out that, for any initial conditions X_0^i, the vectors X_n^i, $i \leq q$, rapidly come close together (due to the communication and convexification). This leads to an (asymptotic in ϵ) jump in the process $X_{[t/\epsilon]}$ at $t = 0$. For this reason, we start $X^\epsilon(\cdot)$ slightly away (n_ϵ steps) from the origin of the $\{X_n\}$ process.

Theorem 3.1. *Assume* (C2.1), *the conditions on* $\{A_n\}$ *in Section* 1, *and* (C3.1), (C3.2), (C3.5) *and either* (C3.3), (C3.4) *or* (C3.3'), (C3.4'). *Then* $X^\epsilon(\cdot)$ *is tight in* $D[0,\infty)$ *and converges weakly to a process* $X(\cdot) = (x(\cdot), \ldots, x(\cdot))$, *where* $x(\cdot)$ *satisfies* (3.1) *with initial condition* x_0, *and* $X(0) = \lim_\epsilon X_0^\epsilon \equiv (x_0, \ldots, x_0)$.

Proof. Part 1. The proofs are essentially the same for the pairs (C3.3), (C3.4) and (C3.3'), (C3.4') and we work only with the first pair. We often use Schwarz' inequality and the inequality (for $a \geq 0$), $E|\Phi(n|k) - \Phi_k|^{1+a} \leq \text{constant}\cdot E|\Phi(n|k) - \Phi_k|$, without specific mention. Iterating (1.1) and letting $n \geq n_\epsilon$ yields

$$\begin{aligned} X_{n+1} &= \Phi(n_\epsilon|0)X_0 + \epsilon \sum_0^{n_\epsilon-1} \Phi(n|k+1)B(X_k,\xi_k) + \epsilon \sum_{n_\epsilon}^{n} \Phi(n|k+1)B(X_k,\xi_k) \\ &= X_0^\epsilon + \epsilon \sum_{n_\epsilon}^{n} \Phi_{k+1}B(X_k,\xi_k) + \epsilon\psi_n^\epsilon + [\Phi(n|0) - \Phi(n_\epsilon|0)]X_0, \end{aligned} \tag{3.2}$$

where

$$\psi_n^\epsilon = \sum_0^n [\Phi(n|k+1) - \Phi_{k+1}]B(X_k,\xi_k).$$

For the purposes of the weak convergence proof, we can assume (w.l.o.g.) that $\{X_k\}$ is bounded by simply truncating the dynamical terms; i.e., changing $B(\cdot,\xi)$ so that it is zero for large $|X|$. If the theorem is true for <u>each such truncation</u>, then by the uniqueness assumption (C3.5), it is true as stated. Henceforth we assume this boundedness.

Part 2.. Next, we show that $\sup_{\epsilon,n} E|\psi_n^\epsilon|^3 < \infty$. All norms are in the ℓ_∞ sense. We have

$$E|\psi_n^\epsilon|^3 \leq \text{constant} \cdot \sum_{i,j,k} E|\Phi(n|i+1) - \Phi_{i+1}||\Phi(n|j+1) - \Phi_{j+1}||\Phi(n|k+1) - \Phi_{k+1}| \cdot$$
$$\cdot [1 + |\tilde{\xi}_i||\tilde{\xi}_j||\tilde{\xi}_k|].$$

By Holder's inequality, the summand is bounded above by

$$E^{1/12}[|\Phi(n|i+1) - \Phi_{i+1}|^{12} \cdot E^{1/12}|\Phi(n|j+1) - \Phi_{j+1}|^{12} \cdot E^{1/12}|\Phi(n|k+1) - \Phi_{k+1}|]^{12} \cdot$$
$$\cdot [1 + E^{3/4}|\tilde{\xi}_i|^4 \cdot E^{3/4}|\tilde{\xi}_j|^4 E^{3/4}|\tilde{\xi}_k|^4].$$

By (C3.2) and the geometric convergence in Lemma 2.1 and the boundedness of $\Phi(n|i)$ and Φ_i, there is a $d \in [0,1)$ such that this term is bounded above by (constant) $d^{n-i}d^{n-j}d^{n-k}$. Thus $\sup_{\epsilon,n} E|\psi_n^\epsilon|^3 < \infty$. From this and (3.2) (and the truncation of $B(\cdot,\cdot)$)

$$\sup_{\epsilon, n \geq n_\epsilon} E|X_{n+1} - X_n|^2/\epsilon^2 < \infty,$$

and $\{|X_{n+1} - X_n|/\epsilon,\ n \geq n_\epsilon,\ \epsilon\}$ is uniformly integrable. Thus, $\{X^\epsilon(\cdot)\}$ is tight in $D[0,\infty)$ and all limit paths are Lipschitz continuous (in t).

Part 3. We fix and work with a weakly convergent subsequence of $\{X^\epsilon(\cdot)\}$, also indexed by ϵ, and with limit denoted by $X(\cdot)$. Skorokod imbedding (see Appendix) will be used where useful, without specific mention. Thus, we can assume, where needed, that $X^\epsilon(\cdot) \to X(\cdot)$ uniformly on bounded time intervals, w.p.1.

We will show, for each real valued function $f(\cdot)$ with compact support and continuous second derivatives, that the $M_f(\cdot)$ defined by

$$(3.3) \qquad M_f(t) = f(X(t)) - f(X(0)) - \int_0^t f_X'(X(s))\bar{\Phi}\,\bar{B}(X(s))ds$$

is a (continuous) martingale. Since $M_f(\cdot)$ is a Lipschitz continuous martingale (since $X(\cdot)$ is Lipschitz continuous), it is a constant. Thus, since $M_f(0) = 0$, we have $M_f(t) = 0$ or, equivalently, $\dot{X} = \bar{\Phi}\,\bar{B}(X)$. By the properties of Φ_k *for each* $i \leq r$, *the* i, $i+r$, ..., $i+qr - r$ *rows of* $\bar{\Phi}$ are equal. Thus all r-vector components of the limit $X(\cdot)$ must be equal, i.e., $X(\cdot)$ is of the form $(x(\cdot), ..., x(\cdot))$, for $x(t) \in E^r$. This and $X = \bar{\Phi}\,\bar{B}(\dot{X})$ implies that $x(\cdot)$ satisfies (3.1).

We need only show the martingale property. To do this, we need only show that for any integer p and continuous bounded $h(\cdot)$ and $t_i \leq t$, $i \leq p$, $s > 0$,

$$(3.4) \qquad Eh(X(t_i), i \leq p)[f(X(t+s)) - f(X(t)) - \int_t^{t+s} f_X'(X(u))\bar{\Phi}\,\bar{B}(X(u))du] = 0.$$

To simplify the notation (and w.l.o.g.), let t and s be integral multiples of $\epsilon m_\epsilon \equiv \delta_\epsilon$ (see (C3.4) for the definition of m_ϵ) and define the index set $I_\ell^\epsilon = \{n: \ell m_\epsilon + n_\epsilon \leq n < \ell m_\epsilon + m_\epsilon + n_\epsilon\}$. By Taylor's Theorem and (3.2),

$$
\begin{aligned}
f(X^\epsilon(t+s)) - f(X^\epsilon(t)) &= \sum_{t\le \ell\delta_\epsilon<t+s}[f(X^\epsilon_{\ell m_\epsilon+m_\epsilon+n_\epsilon}) - f(X^\epsilon_{\ell m_\epsilon+n_\epsilon})] \\
&= \epsilon \sum_{t\le \ell\delta_\epsilon<t+s} f'_X(X^\epsilon_{\ell\delta_\epsilon+n_\epsilon}) \sum_{k\in I^\epsilon_\ell} \Phi_{k+1}B(X_k,\xi_k) + \text{error terms},
\end{aligned}
\tag{3.5}
$$

where the error term is of the order of the sum of (all norms are in the ℓ_∞ sense)

$$
\begin{aligned}
&\epsilon \sum_\ell |\psi^\epsilon_{\ell m_\epsilon+n_\epsilon}|,\ \epsilon^2 \sum_\ell |\psi^\epsilon_{\ell m_\epsilon+n_\epsilon}|^2,\ \epsilon, \\
&\sum_\ell |\Phi(\ell m_\epsilon + m_\epsilon + n_\epsilon|0) - \Phi(n_\epsilon|0)|\cdot|X_0|, \\
&\sum_k \epsilon^2(1 + |\widetilde{\xi}_k|^2),
\end{aligned}
$$

where the sums are over all ℓ such that $t \le \ell\delta_\epsilon < t + s$ and k is summed over $t \le \epsilon k - \epsilon n_\epsilon < t + s$. The mean values of the error terms go to zero as $\epsilon \to 0$.

By (3.5),

$$
\begin{aligned}
&\lim_\epsilon Eh(X^\epsilon(t_i), i \le p)[f(X^\epsilon(t+s)) - f(X^\epsilon(t))] = \\
&\lim_\epsilon Eh(X^\epsilon(t_i), i \le p)\left[\epsilon \sum_{t\le\ell\delta_\epsilon<t+s} f'_X(X_{\ell m_\epsilon+n_\epsilon})\sum_{k\in I^\epsilon_\ell}\Phi_{k+1}B(X_k,\xi_k)\right].
\end{aligned}
\tag{3.6}
$$

We now rearrange the terms in a more convenient way. Define

$$\hat{B}^\epsilon_\ell = \frac{1}{m_\epsilon} \sum_{k\in I^\epsilon_\ell} \Phi_{k+1}B(X_k,\xi_k)$$

and define the function $B^\epsilon(\cdot)$ by

$$B^\epsilon(t) = f'_X(X^\epsilon_{\ell m_\epsilon+n_\epsilon})E_{\ell m_\epsilon+n_\epsilon}\hat{B}^\epsilon_\ell \quad \text{for} \quad \ell\delta_\epsilon \le t < \ell\delta_\epsilon + \delta_\epsilon.$$

Since $X^\epsilon(t_i)$, $i \le p$, is measurable on the σ-algebra $F_{\ell m_\epsilon+n_\epsilon}$, for $\ell\delta_\epsilon \ge t$ (3.6) can be rewritten as

$$\lim_{\epsilon} Eh(X^{\epsilon}(t_i),\ i \leq p)\left[\delta_{\epsilon} \sum_{t \leq \ell\delta_{\epsilon} < t+s} f_X'(X^{\epsilon}_{\ell m_{\epsilon}+n_{\epsilon}})\hat{B}^{\epsilon}_{\ell}\right]$$

(3.7)

$$= \lim_{\epsilon} Eh(X^{\epsilon}(t_i),\ i \leq p)\int_t^{t+s} B^{\epsilon}(u)du.$$

If

(3.8) $$B^{\epsilon}(u) \xrightarrow{\epsilon} f_X'(X(u))\bar{\Phi}\ \bar{B}(X(u))$$

in probability for almost all u, then the second limit in (3.7) would be

$$Eh(X(t_i),\ i \leq p)\int_t^{t+s} f_X'(X(u))\bar{\Phi}\ \bar{B}(X(u))du.$$

Using this and take limits in (3.6) yields the desired result (3.4), and we will be done. Thus, we need only show (3.8).

Fix u and for $\epsilon > 0$, define ℓ_{ϵ} by $u \in [\ell_{\epsilon}\delta_{\epsilon}, \ell_{\epsilon}\delta_{\epsilon} + \delta_{\epsilon})$. Then we need to show that

(3.9) $$\frac{1}{m_{\epsilon}} \sum_{k \in I^{\epsilon}_{\ell_{\epsilon}}} E_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}} f_X'(X_k)\Phi_{k+1}B(X_k,\xi_k) \xrightarrow{P} f_X'(X(u))\bar{\Phi}\ \bar{B}(X(u)).$$

By (C3.2) (and the truncation), we can replace the X_k in (3.9) by $X_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}}$ without changing the limit. Using this and the independence assumption (C3.1), we can rewrite (3.9) as

(3.10) $$\frac{1}{m_{\epsilon}} \sum_{k \in I^{\epsilon}_{\ell_{\epsilon}}} f_X'(X_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}})E_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}}\Phi_{k+1}E_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}}B(X_{\ell_{\epsilon} m_{\epsilon}+n_{\epsilon}},\xi_k)$$
$$+ \text{ error term},$$

where the error term goes to zero in the mean. By the convergence of $X^{\epsilon}(\cdot)$ to $X(\cdot)$ and using $\ell_{\epsilon}\delta_{\epsilon} \to u$, and (C3.3), (C3.4), we get that (3.10) converges in the mean to the right side of (3.9) as $\epsilon \to 0$ and the proof is concluded. (The 'intermediate' details in the last part of the proof are very similar to those in the 'centralized' case. See [8, Chapter 5.2] or [9].). Q.E.D.

State Dependent $\{A_n\}$ and $\{\xi_n\}$ and/or Discontinuous Dynamics. The state dependent 'communication' and noise is most conveniently modeled by a 'Markov' dependence. This will allow $\{A_n\}$ and $\{\xi_n\}$ to depend on the state in a variety of ways: A_n can depend (statistically) on recent events or on changes in the X_n-sequence greater than a given magnitude over some time interval, or on time elapsed since recent communications or on the 'levels' of recent communications (i.e., the degree of 'convexification' or incorporation of received data into ones own estimate can depend on the nature of or timing of recent receptions, transmissions, etc.). To be precise, we suppose that there is a bounded sequence of random variables $\{\tilde{A}_n\}$ such that A_n is a component of $\tilde{A}_n$ and, for each $\epsilon > 0$, $(X_n, \tilde{A}_{n-1}, \xi_{n-1})$ is a Markov process with a homogeneous transition function. The $\tilde{A}_n$ can incorporate other data; e.g., time elapsed since last reception, transmission, etc. The case where some components of $B(\cdot,\cdot)$ are merely indicator functions (hence, not continuous functions) is of particular importance in applications. Such 'Markovianizations' seem to be quite natural for many problems. It might be hard to explicitly evaluate the ODE's here, but the character of the results is clear and precisely what is wanted.

Example. For one example of the appearance of state dependent noise, see the 'routing' problem in [5]. In that example, inputs to a service or communication system occur at random, and the service times are random (correlated or not). The parameter x (the state) determines the probability that incoming events are routed along particular channels. The effective noise is a consequence of the queue length or occupancy level of each channel; it's statistics are dependent on the routing parameter. A Markov dependence model was appropriate there. The routing parameter at time n increased or decreased by ϵ -- depending on whether or not certain events occured at time *n*; hence the dynamics were discontinuous. The model used in this section includes 'decentralized' generalizations of such problems.

Assume that the marginal one-step transition function is of the product (conditionally independent) form, for some P_C and P_N

$$(3.11) \qquad P\{\tilde{A}_1 \in B_1, \xi_1 \in B_2 | X_1, \tilde{A}_0, \xi_0\} = P_C\{\tilde{A}_1 \in B_1 | X_1, \tilde{A}_0\} P_N\{\xi_1 \in B_1 | X_1, \xi_0\},$$

(C denotes 'communication', N denotes 'noise'). The P_C and P_N will not depend on ϵ. *We can allow some ϵ-dependence -- but, in many applications,* ϵ is merely a step size parameter and does not affect the distribution of the A_n, $\widehat{A}_n$ or ξ_n other than via the values of the X_n (e.g., as in the above example). The product from (3.11) is a natural generalization of (C3.1). Here the noise and intercommunication intervals are independent, conditional on the state. For each fixed X, the P_C and P_N in (3.11) can be considered to be one-step transition functions for 'fixed X' Markov chains which we denote by $\{\widehat{A}_n(X)\}$, $\{\xi_n(X)\}$. Let $P_C\{\widehat{A},n,\cdot|X\}$ and $P_N\{\xi,n,\cdot|X\}$ denote the associated n-step transition functions. Then $P_C\{\widehat{A},1,\cdot|X\} = P_C\{\widehat{A}_1 \in \cdot | \widehat{A}_0 = A,X\}$, etc. Let E_C^X and E_N^X denote the associated expectations.

Several assumptions will now be given, followed by some remarks concerning extensions. The assumptions are phrased so as to cover many potential applications.

(C3.6) $E^X[A_n \ldots A_1 | \check{A}_0 = \check{A}] \equiv F_n(\check{A},X)$ is continuous in $(\check{A},X)$.

(C3.7) For each bounded and continuous functions $f_i(\cdot)$, $i = 1.2$, $\int f_1(\xi_1)P_N(\xi,1,d\xi_1|X)$ and $\int f_2(\check{A}_1)P_C(\check{A},1,d\check{A}_1|X)$ are continuous in (ξ,X) and $(\check{A},X)$ respectively.

(C3.8) $\{\xi_n\}$ is bounded.

(C3.9) For each X of the form $X = (x,x,\ldots,x)$, let the pair of processes $\{\check{A}_n(X),\xi_n(X)\}$ associated with the n-step transition function $P_C\{\check{A},n,\cdot|X)P_N\{\xi,n,\cdot|X\}$ have a unique invariant measure and which is of the product form $P^x_C\{\cdot\}P^x_N\{\cdot\}$.

(C3.10) $\int B(X,\xi_1)P\{\xi,1,d\xi_1|X\}$ is continuous in (X,ξ).

Remark. Since the two fixed *X*-processes are independent, the product form in (3.9) will hold if the processes are aperiodic. Under the conditions of Lemma 2.1, the $F_n(\check{A},X)$ in (C3.6) converge geometrically (uniformly in $\check{A},X$) to a function $\Phi(\check{A},X)$, which must be continuous under (C3.6). By the discussion associated with Theorem 3.1, we see that $\Phi(\check{A},X)$ has the form

$$\Phi(\check{A},X) = \begin{bmatrix} \bar{\phi}_1(\check{A},X) & \cdots & \bar{\phi}_q(\check{A},X) \\ \vdots & & \vdots \\ \bar{\phi}_1(\check{A},X) & \cdots & \bar{\phi}_q(\check{A},X) \end{bmatrix} \equiv \begin{bmatrix} \hat{\Phi}(\check{A},X) \\ \vdots \\ \hat{\Phi}(\check{A},X) \end{bmatrix}$$

where $\bar{\phi}_i(\check{A},X)$ is a diagonal $(r\times r)$ matrix. Write $\bar{\phi}_i(\check{A},X) = \text{diag}[\bar{\phi}_{i1}(\check{A},X), \ldots, \bar{\phi}_{ir}(\check{A},X)]$.

If X *takes the form* $X = (x,x,...,x)$ for $x \in E^r$, we simply might write x for X.

(C3.11) *The* ODE (3.12) *has a unique solution for each initial condition* (analogous to (3.1) -- the $\overset{\frown}{A}$ and $\overset{\frown}{\xi}$ are simply averaged out with respect to the invariant measure)

$$
(3.12) \qquad
\left.
\begin{array}{l}
\dot{x}_1 = \sum_j \int \bar{\phi}_{j1}(\overset{\frown}{A},x)P_C^x(d\overset{\frown}{A}\} \int b_1^j(x,\xi^j)P_N^x(d\xi^j) \\
\qquad\qquad \vdots \\
\dot{x}_r = \sum_j \int \bar{\phi}_{jr}(\overset{\frown}{A},x)P_C^x(d\overset{\frown}{A}\} \int b_r^j(x,\xi^j)P_N^x(d\xi^j)
\end{array}
\right. = \hat{\Phi}(x(u)\bar{B}(x(u))
$$

where

$$\hat{\Phi}(x) = \int \hat{\Phi}(\overset{\frown}{A},x)P_C^x(d\overset{\frown}{A}), \quad \bar{B}(x) = \int B(x,\xi)P_N^x(d\xi).$$

Write

$$\bar{\Phi}(x) = \begin{bmatrix} \hat{\Phi}(x) \\ \cdot \\ \cdot \\ \cdot \\ \hat{\Phi}(x) \end{bmatrix}.$$

Under (C3.7), (C3.9) and (C3.10), the right side of (3.12) is continuous.

Remarks on the Assumptions. In many applications, $\overset{\frown}{A}$ takes only a finite number of values. Then the appropriate topology is the discrete topology and the $\overset{\frown}{A}$-continuity required in (C3.6) and (C3.7) always holds -- since then all functions of $\overset{\frown}{A}$ are continuous. The 1-step smoothing assumption in (C3.10) can be replaced by a k-step smoothing assumption -- and Theorem 3.2 will still hold. Since $\Phi(\overset{\frown}{A},X)$ is continuous (see above remark), a Φ_k-analog of (C3.10) is not needed. In (3.12) *we are simply averaging the dynamics with respect to the invariant measures*. If the invariant measure is not unique, then the right side of (3.12) is set valued and P_N^x and P_C^x range over all the invariant measures. We use (C3.8) here to avoid some details. Extensions to cover typical unbounded $\{\xi_n\}$ cases are possible via essentially the same method. To see how this might be done, see the proof for the 'centralized' case in [8, Chapter 5.3] or in [9].

Theorem 3.2. *Assume* (C2.1), *the conditions on* $\{A_n\}$ *in Section* 1, (C3.2) (*without the* $\breve{\xi}$ *component*) *and* (C3.6) *to* (C3.10). *Then* $\{X^\epsilon(\cdot)\}$ *is tight in* $D[0,\infty)$ *and converges weakly to* $X(\cdot) = (x(\cdot), ..., x(\cdot))$, *where* $x(\cdot)$ *satisfies* (3.12) *and* $X(0) = (x(0), ..., x(0)) = \Phi_0 X_0$.

Proof. $\{X^\epsilon(\cdot)\}$ is tight and all limits are Lipschitz continuous for the same reasons as in Theorem 3.1. Let ϵ index a weakly convergent subsequence with limit denoted by $X(\cdot)$. As in Theorem 3.1, $X(\cdot)$ has the form $X(\cdot) = (x(\cdot), ..., x(\cdot))$. Owing to the Markov assumption, E_k denotes conditioning on $(X_k, \xi_{k-1}, \breve{A}_{k-1})$. By the method of proof of Theorem 3.1, we need only show that the left side of (3.9) converges in probability to $f_X'(X(u))\bar{\Phi}(X(u))\bar{B}(X(u))$ for $X(u)$ of the form $X(u) = (x(u), ..., x(u))$. The f_X term does not play an important role and we discard it henceforth.

We use the 'truncation' method and notation discussed in Theorem 3.1. Thus, we can suppose that $B(\cdot,\cdot)$ and $\{X_n\}$ are bounded. For each ν, rewrite (3.9) as (using the conditional independence implied by (3.11))

$$
\begin{aligned}
H^{\epsilon} &\equiv \frac{1}{m_{\epsilon}} \sum_{k \in I^{\epsilon}_{\ell_{\epsilon}}} E_{m_{\epsilon}\ell_{\epsilon}+n_{\epsilon}} E_k \Phi_{k+1} E_k B(X_k,\xi_k) \\
&= \frac{1}{m_{\epsilon}} \sum_{k \in I^{\epsilon}_{\ell_{\epsilon}}} E_{m_{\epsilon}\ell_{\epsilon}+n_{\epsilon}} E_k (A_{k+\nu} \cdots A_{k+1})\, E_k B(X_k,\xi_k) + Q^{\epsilon}_{\nu}
\end{aligned}
\tag{3.13}
$$

where $E|Q^{\epsilon}_{\nu}| \to 0$ uniformly in ϵ, as $\nu \to \infty$, by Lemma 2.1.

We next estimate $E_{k+1}A_{k+\nu} \cdots A_{k+1}$. All norms are in the ℓ_{∞} sense. Since the $\{X_n\}$ and $\{\widehat{A}_n\}$ lie in a compact set, the function of δX defined by

$$\delta_i(|\delta X|) = \sup_{\bar{A},X} |F_i(\widehat{A},X) - F_i(\widehat{A},X + \delta X)|$$

can be supposed to go to zero as $|\delta X| \to 0$. We have $E_n A_n = F_1(\widehat{A}_{n-1},X_n) = F_1(\widehat{A}_{n-1},X_{n-1}) + \Delta_1(\widehat{A}_{n-1},X_{n-1},X_n)$ where $|\Delta_1(\widehat{A}_{n-1},X_{n-1},X_n)| \leq \delta_1(|X_n - X_{n-1}|)$. Next we can write $E_{n-1}A_n A_{n-1} = E_{n-1}(E_n A_n)A_{n-1} = E_{n-1}F_1(\widehat{A}_{n-1},X_{n-1})A_{n-1} + \Delta_1(\widehat{A}_{n-1},X_{n-1},X_n)A_{n-1}$. Note that

$$E_{n-1}F_1(\widehat{A}_{n-1},X_{n-1})A_{n-1} = F_2(\widehat{A}_{n-2},X_{n-1}) = E_C^{X_{n-1}} A_n A_{n-1}, \tag{3.14}$$

which is just the expectation for the 2-step fixed X-process with X fixed at X_{n-1}. Using this and $|A_n| = 1$, we have

$$E_{n-1}A_n A_{n-1} = F_2(\widehat{A}_{n-2},X_{n-2}) + \text{error terms}, \tag{3.15}$$

where

$$
\begin{aligned}
|\text{error terms}| &= |(F_2(\widehat{A}_{n-2},X_{n-1}) - F_2(\widehat{A}_{n-2},X_{n-2})) + \Delta_1(\widehat{A}_{n-1},X_{n-1},X_n)A_{n-1}| \\
&\leq \delta_2(|X_{n-1} - X_{n-2}|) + \delta_1(|X_n - X_{n-1}|)\ .
\end{aligned}
$$

Continuing in this way, we get

$$E_{k+1}A_{k+\nu}\cdots A_{k+1} = F_\nu(\hat{A}_k, X_k) + T_k^{\epsilon,\nu} \tag{3.15}$$

where

$$|T_k^{\epsilon,\nu}| \leq \sum_{i=1}^{\nu} \delta_i(|X_{\nu+k-i+1} - X_{\nu+k-i}|)$$

and $E|T_k^{\epsilon,\nu}| \overset{\epsilon}{\to} 0$ for each ν, uniformly in k, owing to the convergence of the $X^\epsilon(\cdot)$.

Putting the estimate (3.i6) into (3.13) yields that

$$\begin{aligned} H^\epsilon = {} & \frac{1}{m_\epsilon} \sum_{k\in I^\epsilon_{\ell_\epsilon}} E_{m_\epsilon \ell_\epsilon + n_\epsilon} F_\nu(\hat{A}_k, X_k) B(X_k, \xi_k) \\ & + Q^\epsilon_\nu + \frac{1}{m_\epsilon} \sum_{k\in I^\epsilon_{\ell_\epsilon}} E_{m_\epsilon \ell_\epsilon + n_\epsilon} T_k^{\epsilon,\nu} B(X_k, \xi_k). \end{aligned} \tag{3.17}$$

The last two right hand terms in (3.17) go to zero in mean as $\epsilon \to 0$ and then $\nu \to \infty$, and can be neglected. The sequence $F_\nu(\hat{A}, X)$ converges uniformly to the continuous function $\Phi(\hat{A}, X)$ as $\nu \to \infty$. Thus, the limit (as $\epsilon \to 0$, $\nu \to \infty$) of the first term on the right side of (3.16) is the same if $\Phi(\hat{A}, X)$ replaces $F_\nu(\hat{A}, X)$.

Now, we are in a position to use the result of [8, Chapter 5.3] or [9]. By the arguments (for the Markov model) in either of these references (which, when adapted to our current situation, requires the continuity of $\Phi(\cdot,\cdot)$, and (C3.7), (C3.8), (C3.10)) and the fact that $X^\epsilon(\cdot) \to X(\cdot)$, we have

$$\begin{aligned} & \frac{1}{m_\epsilon} \sum_{k\in I^\epsilon_{\ell_\epsilon}} E_{m_\epsilon \ell_\epsilon + n_\epsilon} \Phi(\hat{A}_k, X_k) B(X_k, \xi_k) \\ & \quad \to \int \Phi(\hat{A}, X(u)) B(X(u), \xi) m^{x(u)} (d\hat{A}\, d\xi), \end{aligned} \tag{3.18}$$

where $m^X(\cdot)$ is an invariant measure for the process $\{\hat{A}_n(X), \xi_n(X)\}$. Since $X(u) = (x(u), \ldots, x(u))$, the uniqueness and product form of the invariant measure in (C3.9) yields that $m^X(d\hat{A}d\xi) = P_C^x(d\hat{A}) \cdot P_N^x(d\xi)$. Thus the right side of (3.18) equals

$$\bar{\Phi}(x(u))\bar{B}(x(u)) = \int \Phi(\tilde{A}, x(u)) P_C^{x(u)}(d\tilde{A}) \cdot \int B(x(u), \xi) P_N^{x(u)}(d\xi),$$

and the proof is concluded. Q.E.D.

4. State Space Constraints: A Projection Algorithm

In many applications, it is desired to confine the iterates to a compact set L, and if they ever leave L, the algorithm will project them back onto L. Such algorithms are ubiquitous in applications, even if not explicitly defined or assumed; e.g., the ambiguous notion of 'monitoring' in adaptive control which implicitly assumes some sort of projection. We treat two special but useful, cases.

Assumptions and Problem Formulation

(C4.1) Let $g_i(x)$, $i \leq \alpha$, be real valued continuously differentiable functions on E^r and define $L = \{x: g_i(x) \leq 0, i \leq \alpha\}$. Let L be bounded, convex, and the closure of its interior. Also (w.l.o.g) assume that the gradient $g_{ix}(x)$ is not zero if $g_i(x) = 0$.

Let $\pi_L(y)$ denote the (unique) closest point on L to $y \in E^r$. We use the projected form of algorithm (1.1):

$$\check{X}_{n+1} = A_n X_n + \epsilon b(X_n, \xi_n)$$

(4.1)

$$X^i_{n+1} = \pi_L(\check{X}^i_{n+1}), \quad i \leq q.$$

Thus, each processor projects independently and the constraint set is the same for each. We now set the problem up so that previous results can be used.

Define $\rho_n = (\rho^1_n, ..., \rho^q_n)$, where $\rho^i_n = [X^i_{n+1} - \check{X}^i_{n+1}]/\epsilon$ and define $\check{\psi}^\epsilon_n = \psi^\epsilon_n + \Sigma^n_0[\Phi(n|k+1) - \Phi_{k+1}]\rho_k$. Then for $n \geq n_\epsilon$ (n_ϵ was defined below (C3.4'))

$$X_{n+1} = X_0^{\epsilon} + \epsilon \sum_{n_\epsilon}^{n} \Phi_{k+1} B(X_k, \xi_k) + \epsilon \sum_{n_\epsilon}^{n} \Phi_{k+1} \rho_k + \epsilon \tilde{\psi}_n^{\epsilon}, \tag{4.2}$$

$$+ [\Phi(n|0) - \Phi(n_\epsilon|0)] X_0,$$

where

$$X_0^{\epsilon} = \Phi(n_\epsilon|0)\, X_0 + \epsilon \sum_{0}^{n_\epsilon - 1} \Phi_{k+1}[B(X_k, \xi_k) + \rho_k].$$

The two cases which we treat are covered by the two following assumptions.

(C4.2) *The matrices* $a_{ij}(n)$ *in* A_n *take the form* $a_{ij}(n) = \alpha_{ij}(n) I_r$ *where* $\alpha_{ij}(n)$ *is a scalar valued random variable and* $\Sigma_i \alpha_{ij}(n) = 1$.

Under (C4.2) each of the scalar components 'communicated' from a processor j to processor i are incorporated the same way into the updated estimates of processor i.

(C4.2') *There are bounded* g_{1i} *and* g_{2i} *such that* $L = \{x: g_{1i} \leq x_i \leq g_{2i}, i \leq r\}$.

Definitions. For a vector field $h(\cdot)$ in E^r, define the *projection onto L* by (for $x \in L$) $\pi(x,h(x)) = \lim_{\Delta \to 0} [\pi_L(x + \Delta h(x)) - x]/\Delta$. By the convexity of L, the limit is unique. Define the convex cone

$$C(x) = \left\{ y: y = \sum_{i \in A(x)} \lambda_i g_{ix}(x),\ \lambda_i \geq 0 \right\},$$

where $A(x)$ is the set of constraints $\{i: g_i(x) = 0\}$ (the active constraints at x). Note that $\rho_n^i \in -C(X_{n+1}^i)$. Write $A_n \rho_n = (Z_n^1, \ldots, Z_n^q)$, where $Z_n^i \in E^r$. Under (C4.2), each Z_n^i is a convex combination of vectors in the $-C(X_{n+1}^j)$, $j \leq q$. We will see below that the same property holds under (C4.2'). Similarly for A_k or Φ_k replacing A_n.

The theorem is stated under the conditions of Theorem 3.1, but there is an analogous result under the conditions of Theorem 3.2.

Theorem 4.1. *Assume the conditions of Theorem* 3.1, (C4.1) *and either* (C4.2) *or* (C4.2'). *Let the solution to* (4.3) (*the projected form of* (3.1)) *be unique. Then* $\{X^{\epsilon}(\cdot)\}$ *converges weakly to* $X(\cdot)$, *where* $X(\cdot) = (x(\cdot), ..., x(\cdot))$ *and*

$$\dot{x} = \pi(x,\hat{\Phi}\bar{B}(x)). \tag{4.3}$$

Equivalently

$$\dot{x} = \hat{\Phi}\bar{B}(x) + \nu(x), \tag{4.4}$$

where $\nu(x(t)) \in -C(x(t))$ (*for almost all* t). *Also* $X(0) = \Phi_0 X_0 = (x(0), ..., x(0))$, *if* $X_0^i \in L$.

Proof. Only (4.4) will be proved, since (4.4) implies (4.3). No truncation (see Theorem 3.1) is needed here since $X_n^i \in L$, a compact set. Define the process $\bar{R}^{\epsilon}(\cdot)$ by

$$\bar{R}^{\epsilon}(t) = \epsilon \sum_{n_\epsilon}^{n} \Phi_{k+1}\rho_k \quad \text{for} \quad t \in [(n-n_\epsilon)\epsilon, (n-n_\epsilon+1)\epsilon)$$

(analogous to the definition of $X^{\epsilon}(\cdot)$ above Theorem 3.1). All norms below are in the ℓ_∞ sense. For $X_n^i \in L$, the q r-vector components of $A_n X_n$ are all in L under either (C4.2) or (C4.2'). Thus $|\rho_n^{\epsilon}| \leq |B(X_n,\xi_n)|$. Hence, the proof of uniform integrability of $\{\hat{\psi}_n^{\epsilon}\}$ and $\{\rho_N^{\epsilon}\}$ is the same as that for $\{\psi_n^{\epsilon}\}$ given in Theorem 3.1. Thus $\{X^{\epsilon}(\cdot),\bar{R}^{\epsilon}(\cdot)\}$ is tight and all weak limits are Lipschitz continuous. Henceforth, we fix and work with a weakly convergent subsequence, also indexed by ϵ, and with limit $(X(\cdot),\bar{R}(\cdot))$.

As in Theorem 3.1, for $i \leq q$, the $i, i+r, \ldots, i+rq-r$ rows of Φ_k are equal. Then so are the same components of $\Phi_{k+1}B(X_k,\xi_k)$ and of $\Phi_{k+1}\rho_k$. Thus (as in Theorem 3.1) $X(\cdot) = (x(\cdot), \ldots, x(\cdot))$ and $\bar{R}(\cdot) = (R(\cdot), \ldots, R(\cdot))$, where $x(t)$ and $R(t)$ are in E^r, and

$$\dot{x} = \hat{\Phi}\bar{B}(x) + \dot{R}(t). \tag{4.5}$$

Obviously $x(t) \in L$. Thus, we need only show that $\dot{R}(t) \in -C(x(t))$ for almost all t.

Write $X^\epsilon(\cdot) = (X^{\epsilon,1}(\cdot), \ldots, X^{\epsilon,q}(\cdot))$. Let $x(t)$ be in the <u>interior</u> of L for $t \in [t_1,t_2]$ with $t_1 < t_2$. Then, by the weak convergence (i.e., convergence of all $X^{\epsilon,i}(\cdot)$ to $x(\cdot)$) the $X^{\epsilon,i}(t)$, $i \leq q$, are strictly interior to L on $[t_1,t_2]$ with a probbility which tends to unity as $\epsilon \to 0$. Thus, for small ϵ, the cones $C(X^{\epsilon,i}(t))$, $i \leq q$, $t_1 \leq t \leq t_2$, will be empty with a probability which tends to unity as $\epsilon \to 0$. Thus $R(t) = 0$ for $t_1 \leq t \leq t_2$.

We need now only consider the case where $x(t)$ is on the boundary of L for $t \in [t_1,t_2]$, $t_1 < t_2$. Skorohod imbedding will be used (see Appendix), so that we can assume that the convergence is with probability one on each bounded time interval. Note that $C(x)$ is an upper semicontinuous function of x in the sense that if $x_n \to x$, then

$$C(x) \supset \bigcap_n \bigcup_{k=n}^{\infty} C(x_n). \tag{4.6}$$

Let $(g_{i_1 x}(x(t)), \ldots, g_{i_a x}(x(t)) \equiv (\nu_1, \ldots, \nu_a)$ be the gradient vectors of the active constraints at $x = x(t)$, and let C_β denote the convex cone formed by the vectors in a β-neighborhood of $(\nu_1, \ldots, \nu_a)$.

By the weak convergence (i.e., the convergence of all $X^{\epsilon,i}(\cdot)$ to $x(\cdot)$) and (4.6), for each $\beta > 0$ and $\gamma > 0$, there are $\beta_1 > 0$ and $\epsilon_1 > 0$ such that for $\epsilon \leq \epsilon_1$,

(4.7) $$P\{\rho_k^i \in -C_\beta,\ i \leq q, \text{ all } k \text{ such that } |\epsilon(k-n_\epsilon)-t| \leq \beta_1\} \geq 1 - \gamma;$$

i.e., for $\epsilon(k-n_\epsilon)$ close enough to t, the ρ_k^i are in a 'small neighborhood' of $-C(x(t))$ with a probability close to unity.

Now, assume (C4.2). Then, each of the q r-vector components of $\Phi_{k+1}\rho_k$ is also in such a 'small neighborhood' with a probability close to unity, for $\epsilon(k-n_\epsilon)$ close to t. This implies that $R(t) \in -C(x(t))$, for almost all t.

Write $x(t) = (x_1(t), \ldots, x_r(t))$. Assume (C4.2'), and let $\epsilon(k-n_\epsilon)$ be close to t. Then $C(x(t))$ is particularly simple. Write $\rho_k^j = (\rho_k^{j1}, \ldots, \rho_k^{jr})$ where the ρ_k^{ji} are scalar valued. If $x_i(t) = g_{1i}$ (the lower limit) then (using the weak convergence) $X^\epsilon(\cdot) \Rightarrow (x(\cdot), \ldots, x(\cdot))$, the ρ_k^{ji} must be (asymptotically in ϵ) ≥ 0 for all j, with a probability arbitrarily close to unity. Similarly, if $x_i(t) = g_{2i}$ (the upper limit), then (asymptotically in ϵ) the ρ_k^{ji} must be ≤ 0 for all j. By the properties of Φ_{k+1}, the same property must hold for the respective components (i, i+r, i+2r, ...) of $\Phi_{k+1}\rho_k$.

The conclusion follows from this last remark, since if $x = (x_1, \ldots, x_r)$, where $x_i = g_{1i}$, $i \leq r_1$, $x_i = g_{2i}$, $r_1 < i \leq r_2$ and $g_{1i} < x_i < g_{2i}$, $r_2 < i \leq r$, then we have that $-C(x)$ is the collection of vectors whose first r_1 components are nonnegative, the next $r_2 - r_1$ are nonpositive and the last $r - r_2$ are zero. Q.E.D.

5. The Asymptotics of $X^\epsilon(\cdot)$ for Large t and Small ϵ

Weak convergence in $D[0,\infty)$ or in $C[0,\infty)$ basically gives information on the locations and/or distribution of $X^\epsilon(\cdot)$ for small ϵ, and for t confined to some large -- but still bounded interval. See, e.g., the discussion of the topology of these spaces in the Appendix. It is important to have a convergence result which is valid uniformly in (large) t for small ϵ, and such a result is readily available by appropriate modifications of the previous results. One usually requires that the ODE satisfied by the limit processes is stable, hence we assume

(C5.1) Let (3.1) (or (3.12) for the state dependent $\{A_n,\xi_n\}$ case) have a unique stable (in the sense of Liapunov) point θ which is globally attracting.

Let $t_\epsilon \to \infty$ as $\epsilon \to 0$. Quite generally, if (C5.1) (and the conditions of Theorems 3.1 or 3.2) holds, then $X^\epsilon(t_\epsilon + \cdot)$ converges weakly to a constant process $\bar{X}(\cdot)$, where $\bar{X}(t) = (\theta, ..., \theta)$. This is precisely the desired asymptotic result, since it says (roughly) that if the algorithm is 'stable' then, after a fixed 'transient period' (independent of ϵ), the $X^{i\epsilon}(\cdot)$ are arbitrarily close to θ in the sense of weak convergence.

Discussion of the Main Idea of the Development. Suppose that the set

(5.1) $$M = \{X^\epsilon(t), t \geq 0, \epsilon > 0\}$$

is bounded in probability (tight); i.e., for each $\eta > 0$ there is a $k_\eta < \infty$ such that $P\{|X^\epsilon(t)| \geq k_\eta\} \leq \eta$, for all $\epsilon > 0$, $t \geq 0$. Then it is easy to show that $X^\epsilon(t_\epsilon + \cdot) \Rightarrow \bar{X}(\cdot)$. To see this, choose $T > 0$ and consider a convergent subsequence of the pair of processes $\{X^\epsilon(t_\epsilon + \cdot), X^\epsilon(t_\epsilon - T + \cdot)\}$, with limit denoted by $(X(\cdot),X_T(\cdot)) = (x(\cdot), ..., x(\cdot); x_T(\cdot), ..., x_T(\cdot))$ (recall that all the r-vector components of the limits are equal). We have $X(0) = X_T(T)$. The value of $X_T(0)$ is unknown -- but all the possible such $X_T(0)$, over all T and convergent subsequences, belong to a tight set, with the same η and k_η as above. By this and the stability condition (A5.1) and Theorem 3.1 (or Theorem 3.2), for any $\delta > 0$ there is a $T_\delta < \infty$ such that for $T \geqslant T_\delta$, $X_T(T) = (x_T(T), ..., x_T(T))$ will be in a δ-neighborhood of $(\theta, ..., \theta)$ with probability $\geqslant 1-\delta$. This yields the desired conclusion, since it implies that $X(0) = (\theta, ..., \theta)$ w.p.1. Thus, to get the asymptotic (in t and ϵ) result, only (5.1) must be shown.

Next, consider the projection algorithm of Section 4 and assume (C5.1') in lieu of (C5.1):

(C5.1') Let (4.4) have a unique stable (in the sense of Liapunov) point θ which is attracting in L.

Under (C5.1'), (5.1) is automatically bounded and if $t_\epsilon \to \infty$ as $\epsilon \to 0$ then under the additional conditions of Section 4, $X^\epsilon(t_\epsilon + \cdot) \Rightarrow \bar{X}(\cdot)$, where $\bar{X}(t) = (\theta, ..., \theta)$. Some form of projection algorithm is usually used in practical algorithms, and so the tightness condition on (5.1) is not burdensome.

Sharper Bounds on the Asymptotic Errors $(X_n^i - \theta)$, for Large ϵn and Small ϵ. Under additional 'stability' conditions, one can get order of magnitude estimates for $(X^{i,\epsilon}(t) - \theta)$ for large t and small ϵ. We do one case here in preparation for the rate of convergence work in the next section. We will need:

(C5.2) There is a twice continuously differentiable Liapunov function $0 \leq \bar{V}(x) \to \infty$ and $\bar{V}(x) > 0$ for $x \neq 0$ such that for some $\lambda > 0$ and $K < \infty$, $\bar{V}_x'(x)\hat{\Phi}\bar{B}(x) \leq -\lambda\bar{V}(x)$, $|\bar{V}_x(x)|^2 \leq K|\bar{V}(x) + 1|$ and $\bar{V}_{xx}(\cdot)$ is bounded.

Define

(5.2) $\qquad V(X) = \sum_1^q \bar{V}(X^i)$, for $X = (X^1, \ldots, X^q)$.

(C5.3) (C3.2), but where $B_0(X,\hat{\xi})$ and $B_1(X)$ are bounded and have bounded and continuous X-derivatives (uniformly in $\hat{\xi}$, for B_0).

(C5.4) There is a constant K such that

$$E\left|\sum_{\nu}^{\nu+m} E_\nu(\Phi_{k+1}B(X,\xi_k) - \bar{\Phi}\,\bar{B}(X))\right|^2 \leq K[V(X) + 1],$$

for all positive m and ν. Similarly for the derivatives B_X and $\bar{B}_X$ replacing B and $\bar{B}_1$ respectively.

Remark. (C5.4) essentially implies a 'low' correlation between data in the remote past and in the distant future. There is an analogous result to Theorem 5.1 for the state dependent $\{A_n,\xi_n\}$ case, and for the constrained case.

Theorem 5.1. Assume (C5.1) to (C5.4). There is an $N_\epsilon < \infty$ for each small ϵ such that

(5.3) $\qquad EV(X_n) = 0(\sqrt{\epsilon})$, $n \geq N_\epsilon$.

Proof. We always assume $n \geq n_\epsilon$ so that $E|\Phi(n|0) - \Phi_0|^a = 0(\epsilon^2)$, for any $a > 0$. Write

$$X_{n+1} - X_n = [\Phi(n|0) - \Phi(n-1|0)]X_0 + \epsilon(\psi_n^\epsilon - \psi_{n-1}^\epsilon) \tag{5.4}$$
$$+ \epsilon \bar{\Phi} \bar{B}(X_n) + \epsilon[\Phi_{n+1}B(X_n,\xi_n) - \bar{\Phi} \bar{B}(X_n)]$$

and

$$E_nV(X_{n+1}) - V(X_n) = \epsilon V_X'(X_n) E_n[\Phi(n|0) - \Phi(n-1|0)]X_0$$
$$+ \epsilon V_X'(X_n)E_n(\psi_n^\epsilon - \psi_{n-1}^\epsilon) + \epsilon V_X'(X_n) \bar{\Phi} \bar{B}(X_n) \tag{5.5}$$
$$+ \epsilon V_X'(X_n)E_n[\Phi_{n+1}B(X_n,\xi_n) - \bar{\Phi} \bar{B}(X_n)] + \text{error term}$$

where $E|\text{error term}| = 0(\epsilon^2)$. By (C5.2) and $n \geq n_\epsilon$, the expectation of the first term on the rhs of (5.5) is $0(\epsilon^2)(1 + EV(X_n))$. Write Φ_n in the form

$$\Phi_n = \begin{bmatrix} \hat{\Phi}_n \\ \cdot \\ \cdot \\ \cdot \\ \hat{\Phi}_n \end{bmatrix}$$

where $\hat{\Phi}_n$ is a r × qr matrix. For $n \geq n_\epsilon$,

$$|X_n^i - X_n^j| = 0_n(\epsilon^2) + 0(\epsilon)|\psi_n^\epsilon|, \tag{5.6}$$

where $E|0_n(\epsilon^2)|^2 = 0(\epsilon^4)$, uniformly in $n \geq n_\epsilon$.

Using (5.6), rewrite the last two terms on the right side of (5.5) as, respectively,

$$\epsilon \sum_i \bar{V}_x'(X_n^i)\, \hat{\Phi}\, \bar{B}\,(X_n^i) + \text{error term},$$

(5.7)
$$\epsilon \sum_i \bar{V}_x'\,(X_n^i)E_n \left[\hat{\Phi}_{n+1}B(X_n^i\,,\,\xi_n) - \hat{\Phi}\,\bar{B}(X_n^i)\right] + \text{error term},$$

where by (C5.2) $E|\text{error term}| = 0(\epsilon^2)(1 + EV(X_n))$.

We now define the perturbations to the Liapunov function.

Define

$V_1^\epsilon(n)$ by $V_1^\epsilon(n) = -\epsilon V_X'(X_n)\psi_{n-1}^\epsilon$. We have

(5.8) $$E|V_1^\epsilon(n)| = 0(\epsilon)(1 + EV(X_n))$$

(5.9) $$E\ V_1^\epsilon(n+1) - EV_1^\epsilon(n) \leq -\epsilon\ EV_X'(X_n)(\psi_n^\epsilon - \psi_{n-1}^\epsilon) + 0(\epsilon^2)\ E\ (1 + V(X_n)).$$

Define $V_2^{i,\epsilon}(n)$

(5.10) $$V_2^{i,\epsilon}(n) = \epsilon \sum_{j=n}^{\infty} \bar{V}_x'\,(X_n^i)\ E_n \left[\hat{\Phi}_{j+1}\ B(X_n^i\,,\xi_j) - \hat{\Phi}\ \bar{B}\,(X_n^i)\right].$$

By (C5.2) and (C5.4),

(5.11) $$E|V_2^{i,\epsilon}(n)| = 0(\epsilon)(1 + EV(X_n)).$$

Also,

(5.12) $$E_n V_2^{i,\epsilon}\ (n+1) - V_2^{i,\epsilon}\,(n) = -\epsilon\ \bar{V}_x'\ (X_n^i)\ E_n \left[\hat{\Phi}_{n+1}\ B(X_n^i,\xi_n) - \hat{\Phi}\ \bar{B}(X_n^i)\right]$$
$$+ \text{ error term}\,,$$

where by (C5.4), $E|\text{error term}| = 0(\epsilon^2)(1 + EV(X_n))$.

Now, define the perturbed Liapunov function $V^\epsilon(n) = V(X_n) + V_1^\epsilon(n) + \sum_1^q V_2^{i,\epsilon}(n)$, and evaluate $E_n V^\epsilon(n+1)-V^\epsilon(n)$ and cancel the terms $\pm\epsilon V_x'(X_n)(\psi_n^\epsilon -$

ψ^{ϵ}_{n-1}) and $\pm\epsilon \sum_{j} \bar{V}_x'(X^i_n)E_n[\hat{\Phi}_{n+1}B(X^i_n,\xi_j) - \hat{\Phi}\,\bar{B}(X^i_n)]$ to get

$$E_nV^{\epsilon}(n+1) - V^{\epsilon}(n) = \epsilon \sum_{i} \bar{V}_x'(X^i_n)\,\hat{\Phi}\,\bar{B}\,(X^i_n) + \text{error terms,} \tag{5.13}$$

$$E|\text{error term}| = 0(\epsilon^2)(1 + EV(X_n)).$$

Using (C5.2) and the bounds on $E|V^{\epsilon}_1(n)|$ and on $E|V^{i,\epsilon}_2(n)|$, we get

$$\begin{aligned} EV^{\epsilon}(n+1) - EV^{\epsilon}(n) &\leqslant \lambda\epsilon \sum_{i} E\bar{V}(X^i_n) + 0(\epsilon^2)\,(1 + EV(X_n)) \\ &\leqslant -\lambda\epsilon\; EV^{\epsilon}(n) + 0(\epsilon^2)\,(1 + EV^{\epsilon}(n)). \end{aligned} \tag{5.14}$$

Hence, for small $\epsilon>0$,

$$EV^{\epsilon}(n) \leqslant (1 - \tfrac{\lambda\epsilon}{2})^{n-n_\epsilon}\; V^{\epsilon}(n_{\epsilon}) + 0(\epsilon). \tag{5.15}$$

This together with the bounds on $E|V^{\epsilon}_1(n)|$ and on $E|V^{i,\epsilon}_2(n)|$ yield the Theorem. Q.E.D.

6. Rate of Convergence: Qualitative Asymptotic Properties

The Liapunov function in (C5.2) is often locally quadratic about θ in the sense that $\bar{V}(x) = x'Qx + 0(|x|^3)$ for $Q > 0$. If this is true, then Theorem 5.1 implies that $\{(X_n^i - \theta)/\epsilon^{1/2}, i \leq q, n \geq N_\epsilon, \epsilon > 0\}$ is tight. In this section, we will suppose that there are $\tilde{N}_\epsilon < \infty$ for each small $\epsilon > 0$ so that

$$(6.1) \qquad \left\{\frac{X_n^i - \theta}{\sqrt{\epsilon}}, \; i \leq q, \; n \geq \tilde{N}_\epsilon, \; \epsilon > 0\right\} \text{ is tight, } Eb^i(\theta,\xi_k^i) \equiv 0.$$

Under (6.1), one can apply the methods of the 'centralized' case to get a classical rate of convergence result.

Much information concerning the asymptotic behavior and comparison with other algorithms can be obtained from such a result. The method and results will be discussed in an informal way so that the main ideas are clear. Despite the informality, the conditions needed for the proof will generally be stated. The proofs follow standard lines in weak convergence theory, and are not hard. Our aim is to exhibit the asymptotic behavior of the suitably normalized errors, then specialize them to simple cases where a comparison can be made with 'centralized' forms of the algorithm, so that one can see the effects and value of the decentralization, and evaluate alternative forms of communication and algorithms. The discussion is continued in the next section. Such insights are needed at this stage of development of the 'decentralized' algorithms, as a guide to future developments and are perhaps more important than a rigorous development along the standard lines. We will use the assumption (6.1), the boundedness of $B(\cdot,\cdot)$ in each bounded X-set and that $B(\cdot,\xi)$ has a continuous (uniformly in ξ) derivative, and $EB(X,\xi) \equiv 0$.

For any R^s valued function $p(\cdot) = (p^1(\cdot), \ldots)$ of x (or X), let $(p(\theta))_x$ denote the (Jacobian) matrix whose ith <u>row</u> is the x (respectively, X) gradient of $p^i(\cdot)$. Recall the definitions of $\bar{\Phi}_i$ and $\hat{\Phi}$ (above (C3.5)), and of $\hat{\Phi}_n$ (in Theorem 5.1). Define the matrix $M = (\hat{\Phi}\,\bar{B}(\theta))_x$ and suppose that it is stable. Let

$$\frac{1}{m} \sum_{m}^{n+m} (\hat{\Phi}_{k+1} B(\theta,\xi_k))_x \to (\hat{\Phi}\,\bar{B}(\theta))_x = M$$

in probability as $n \to \infty$ and $m \to \infty$.

Define $U_n^\epsilon = (X_n - \bar{\theta})/\sqrt{\epsilon}$, where $\bar{\theta} = (\theta, \theta, \ldots, \theta)$. Recall the definition of n_ϵ given below (C3.4') and that n_ϵ can be chosen such that $\sqrt{\epsilon}n_\epsilon \to 0$. Given $N > 0$, we have, for $n \geqslant n_\epsilon + N$,

$$U_{n+1}^\epsilon = \Phi(n|N)U_N^\epsilon + \sqrt{\epsilon} \sum_{N}^{N+n_\epsilon} \Phi(n|k+1)B(X_k,\xi_k)$$

$$+ \sqrt{\epsilon} \sum_{N+n_\epsilon+1}^{n} \Phi_{k+1}B(X_k,\xi_k) + \sqrt{\epsilon}\,\hat{\psi}_n^\epsilon, \tag{6.2}$$

$$\hat{\psi}_n^\epsilon = \sum_{N+n_\epsilon}^{n} [\Phi(n|k+1) - \Phi_{k+1}]B(X_k,\xi_k).$$

Define (for $n \geqslant N + n_\epsilon$)

$$W_n^\epsilon = \sqrt{\epsilon} \sum_{N+n_\epsilon+1}^{n} \Phi_{k+1}B(\theta,\xi_k).$$

Let $N \geqslant \widetilde{N}_\epsilon$. For $t \geqslant 0$, define the process $U^\epsilon(\cdot)$ by $U^\epsilon(t) = U_n^\epsilon$ for $t \in [\epsilon(n-N-n_\epsilon), \epsilon(n-N-n_\epsilon+1))$ and define $W^\epsilon(\cdot)$ similarly from $\{W_n^\epsilon\}$. By Taylor's Theorem and the definition of n_ϵ,

$$
\begin{aligned}
U_{n+1}^\epsilon &= \Phi_N U_N^\epsilon + 0_n(\epsilon^2)U_N^\epsilon + \widetilde{0}_n(n_\epsilon\sqrt{\epsilon}) + 0(\sqrt{\epsilon})\hat{\psi}_n^\epsilon \\
&\quad + \epsilon \sum_{N+n_\epsilon+1}^{n} (\Phi_{k+1}B(\theta,\xi_k))_X U_k^\epsilon + W_n^\epsilon + o(\epsilon) \sum_{N+n_\epsilon+1}^{n} 0(|U_k^\epsilon|),
\end{aligned}
\tag{6.3}
$$

where $E|0_n(\epsilon^2)|^2 = 0(\epsilon^4)$ since $n \geqslant n_\epsilon$, and $E|\widetilde{0}_n(n_\epsilon\sqrt{\epsilon})| = 0(n_\epsilon\sqrt{\epsilon})$. Also $(\Phi_{k+1}B(\theta,\xi_k))_X$ denotes the matrix whose rows are the X-gradients of the components of $\Phi_{k+1}\, B(X,\xi_k)$ evaluated at $X = (\theta,\theta,...)$.

In order to study the weak convergence of $U^\epsilon(\cdot)$, we can truncate the dynamics (as in Theorem 3.1) if $\{U_k^\epsilon\}$ is not bounded: wherever U_k^ϵ appears in (6.3), we simply replace it by $U_k^\epsilon q_m(U_k^\epsilon)$, where $q_m(u) = 1$ for $|u| \leqslant m$, and is a smooth function with compact support. We get the weak convergence with use of q_m, and then let $m \to \infty$. The uniqueness of the solution to the limit equation (6.9) below guarantees that the procedure works. For notational simplicity -- we simply suppose that $\{U_k^\epsilon\}$ is bounded. Suppose that $\{W^\epsilon(\cdot)\}$ is tight and has continuous limits. Then, this also holds for $\{U^\epsilon(\cdot)\}$. Also, the second, third, fourth and last terms on the right side of (6.3) disappear in the limit. The limit of any convergent subsequence satisfies

$$
U(t) = U(0) + \int_0^t (\bar{\Phi}\ \bar{B}(\theta))_x U(s)ds + W(t),
\tag{6.4}
$$

where $W(\cdot)$ is the limit of $\{W^\epsilon(\cdot)\}$.

The Limits of $\{W^\epsilon(\cdot)\}$. Under broad conditions $W(\cdot)$ is a Wiener process with covariance

$$(6.5) \qquad t\sum_{-\infty}^{\infty} E\Phi_{k+1}B(\theta,\xi_k)B'(\theta,\xi_0)\Phi_1',$$

where the expectation in (6.5) is to be interpreted in the ergodic sense:

$$\lim_{m} \frac{1}{m} \sum_{j=n}^{m+n} E\Phi_{k+j+1}B(\theta,\xi_{k+j})B'(\theta,\xi_j)\Phi_{j+1}',$$

We now give some conditions under which $W(\cdot)$ is the asserted Wiener process. Let

$$(6.6) \qquad E\left|\sum_{n}^{n+m-1} \Phi_{k+1}B(\theta,\xi_k)\right|^4 \leq \text{Constant} \cdot m^2.$$

then $\{W^\epsilon(\cdot)\}$ is tight and all limits are continuous [6]. If

$$(6.7) \qquad \sum_{n}^{m+n-1} \Phi_{k+1}B(\theta,\xi_k)/\sqrt{m}$$

converges in distribution to a normal random variable (with mean zero) as $n \to \infty$ and $m \to \infty$, then $W(\cdot)$ is a Gaussian process. If, for $t_1 \leq t_2 \leq t_3 \leq t_4$,

$$(6.8) \qquad E[W^\epsilon(t_4) - W^\epsilon(t_3)][W^\epsilon(t_2) - W^\epsilon(t_1)]' \xrightarrow{\epsilon} 0,$$

then the increments of the limit $W(\cdot)$ are orthogonal and the limit is a (nonstandard) Wiener process. The proofs follow standard lines in weak convergence theory [6]. The properties (6.6), (6.8) hold if the $\{A_k\}$ is independent of the $\{\xi_k\}$ and the dependence among the ξ_k decreases fast enough as the time difference increases. Henceforth we assume that $W(\cdot)$ is the zero mean Wiener process with covariance (6.5).

For the same reasons that the $X(\cdot)$ of Section 3 took the form $X(\cdot) = (x(\cdot), ..., x(\cdot))$ for $x(t) \epsilon E^r$, we have $U(\cdot) = (u(\cdot), ..., u(\cdot))$ and $W(\cdot) = (w(\cdot), ..., w(\cdot))$. Then (6.4) reduces to

$$(6.9) \qquad du = Mu\, dt + dw.$$

The covariance of $w(1)$ can be obtained from (6.5): Writing

$$\Phi_k = \begin{bmatrix} \phi_1(k), & \cdots, & \phi_q(k) \\ \vdots & & \vdots \\ \phi_1(k), & \cdots, & \phi_q(k) \end{bmatrix},$$

where the $\phi_i(k)$ are diagonal, (6.5) reduces to

$$(6.10) \qquad \text{cov } w(1) = \bar{R} = \sum_{-\infty}^{\infty} E\left[\sum_1^q \phi_i(k+1)b^i(\theta,\xi_k^i)\right]\left[\sum_1^q \phi_i(k+1)b^i(\theta,\xi_k^i)\right]'.$$

If $N \to \infty$ fast enough as $\epsilon \to 0$, then the limit $u(\cdot)$ is the stationary solution to (6.9).

The stationary covariance

$$(6.11) \qquad \int_0^\infty e^{Mt}\, \bar{R} e^{M't} dt.$$

of (6.9) is a standard measure of the 'rate of convergence' or asymptotic quality of the algorithm, and can be used as a basis of comparison among alternative algorithms.

A Special Case. We specialize to a simple case in order to get some insight into the asymptotic behavior. Let $\{\xi_k\}$ be independent of $\{A_k\}$ with $\{\xi_k^i, i \leq q, k = 1,2, ...\}$ mutually independent with cov $b^i(\theta,\xi_k^i) \equiv R_i$. Then

$$\text{(6.12)} \qquad \mathrm{cov}\ w(1) = \sum_{1}^{q} \lim_{m} \frac{1}{m} \sum_{n}^{n+m} \phi_i(k) R_i \phi_i(k).$$

A Scalar System. Let $r = 1$. Then $\phi_i(k)$ and $\bar{\phi}_i$ are scalars and

$$\sum_{1}^{q} \bar{\phi}_i = 1 = \sum_{1}^{q} \phi_i(k).$$

Let $b^i(\cdot,\cdot) = b(\cdot,\cdot)$ and $R_i = R$ not depend on i. Then (6.9) becomes ($\bar{b}_x(\theta) < 0$)

$$\text{(6.13)} \qquad du = \bar{b}_x(\theta) u\ dt + \sigma_D d\breve{w},$$

where $\breve{w}(\cdot)$ is a standard Wiener process and (where by the expectation E, we mean the ergodic mean in (6.12))

$$\sigma_D^2 = R \sum_{1}^{q} E\phi_i^2(n).$$

The stationary variance of $u(\cdot)$ is $\sigma_D^2/2|\bar{b}_x(\theta)| \equiv \mathrm{var}_D$.

Comparison with a 'Centralized' Algorithm. Define the following centralized algorithm, under the scalar system assumptions of the above paragraph

$$\text{(6.14)} \qquad Z_{n+1} = Z_n + \epsilon b(Z_n, \xi_n^1), \quad \{\xi_n^1,\ n = 1,2, ...\} \text{ i.i.d.}$$

Define $V_n = (Z_n - \theta)/\sqrt{\epsilon}$ and define $v^\epsilon(\cdot)$ by $v^\epsilon(t) = V_n$ on $[n\epsilon, n\epsilon+\epsilon)$. If $t_\epsilon \to \infty$ fast enough as $\epsilon \to 0$, then under appropriate conditions [8] $v^\epsilon(t_\epsilon + \cdot) \Rightarrow v(\cdot)$ where

$$\text{(6.15)} \qquad dv = \bar{b}_x(\theta) v\ dt + \sqrt{R}\, d\breve{w}.$$

The stationary variance of (6.15) is $R/2|\bar{b}_x(\theta)| = \mathrm{var}_C$. Since

$$\frac{\mathrm{var}_D}{\mathrm{var}_C} = E \sum_1^q \phi_i^2(n) < 1,$$

the decentralized algorithm yields an improvement. The infima of the ratio occurs when the $E\phi_i^2(n)$, $i \leq q$, are all equal, an unattainable case (to which we can come close -- see Section 7). In this limit, $1/q = \mathrm{var}_D/\mathrm{var}_C$.

A fairer comparison accounts for the fact that the decentralized algorithm uses a total of q observations per iterate. Using the same number in the centralized algorithm (6.14) we rewrite it as

(6.16) $$\bar{Z}_{n+1} = \bar{Z}_n + \frac{\epsilon}{q}\sum_1^q b(\bar{Z}_n,\xi_n^i), \quad \{\xi_n^i,\ i \leq q,\ n = 1,2,\ldots\} \text{ i.i.d.}$$

Define $\bar{V}_n^\epsilon$ and $\bar{v}^\epsilon(\cdot)$ as the V_n^ϵ and $v^\epsilon(\cdot)$ were defined, but based on $\{\bar{Z}_n\}$. Under appropriate conditions $\bar{v}^\epsilon(t_\epsilon+\cdot) \Rightarrow \bar{v}(\cdot)$ where

(6.17) $$d\bar{v} = \bar{b}_x(\theta)\bar{v}\,dt + \sqrt{R/q}\,d\check{w},$$

with stationary covariance $R/2q|b_x(\theta)| = \mathrm{var}_{qC}$ and

(6.18) $$\mathrm{var}_D/\mathrm{var}_{qC} = q\sum_1^q E\Phi_i^2(n) \geq 1.$$

The ratio (6.18) can be used to decide on the proper tradeoff between the asymptotic error and the communication policy. Reasons why the decentralized algorithm might be preferable are discussed in the next section. Analogous results are, of course, obtainable for the general vector case.

7. Asymptotic Properties: Discussion and Comparison

Independent $\{A_n\}$. We evaluate $\mathrm{var}_D/\mathrm{var}_C$ under the conditions of the last subsection of Section 6, where $q = 2$ and the $\{A_n\}$ are i.i.d. In particular, let $c \in [0,1)$, and let the processors act independently, with p = probability i communicates to $j \neq i$ at time n. With no communication (probability $(1-p)^2$), $A_n = I$; if 2 communicates to 1 -- but not conversely (probability $p(1-p)$), then

$$A_n = \begin{bmatrix} 1-c & c \\ 0 & 1 \end{bmatrix} = A^{21},$$

If 1 communicates to 2 (but not conversely), then

$$A_n = \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} = A^{12};$$

If both communicate to each other, then

$$A_n = \begin{bmatrix} 1-c & c \\ c & 1-c \end{bmatrix} = A^{0}.$$

Refer to Table 1. The <u>optimum</u> value of the ratio of the variances is unity, a value closely approximated by small c. Clearly a larger p is desirable. As $c \to 0$, the ratio improves -- but the size of the ψ_n^ϵ would increase. This implies that one must wait longer for the stationary variance to be a good indicator of the actual performance (the effects of the communication are realized more slowly). Similarly, for small p. But, in all cases, the average performance is much better than that for the centralized algorithm (6.14).

c p	.05	.25	.5
.1	1.036	1.13	1.312
.3	1.016	1.10	1.26
.7	1.008	1.04	1.13

Table 1. Values for $2\,\text{var}_D/\text{var}_C = \text{var}_D/\text{var}_{2C}$

A Deterministic Communication Scheme. We retain the assumptions of the last subsection, except for those on the communications. Let m and m_1 be integers with $m_1 \leq m/2$. Processor 2 communicates to 1 each m units of time, and 1 communicates to 2 m_1 units of time later. We use A^{12}, A^{21} (when $m_1 \neq 0$) and A^0 (when $m_1 = 0$). For $m_1 = 0$, $(2\text{var}_D/\text{var}_C) = 1$, for all $0 < 1 < c$. For $m_1 \neq 0$, we have Table 2.

c	$2\,\text{var}_D/\text{var}_C$
.1	1.0028
.3	1.03
.5	1.11

Table 2. $2\,\text{var}_D/\text{var}_C = \text{var}_D/\text{var}_{2C}$

The values of m and m_1 appear only in the values of ψ_n^ϵ, which increases as m and m_1 increase. The values of $\text{var}_D/\text{var}_{2C}$ are substantially worse when processor 1 communicates to processor 2 more often than the reverse communication rate, for deterministic communication times. This suggests that a relatively balanced communication strategy is better and that a processor should 'respond' as soon as possible after it receives a 'message' from another processor.

Discussion. It is clear that the decentralized algorithm takes advantage of the possibilities of parallel processing, since its variance is better than that of the classical algorithm (6.14), and can be nearly as good as that of the fully centralized algorithm (6.16). But there is another advantage -- which can be considerable. Simulations with recursive algorithms such as (6.14) indicate that a key problem concerns the frequently slow recovery from the effects of large 'bursts' of noise; i.e., from a large 'random' jump in the state value. This effect would not show up in the asymptotic variance estimates, but is of considerable importance in practice, particularly when the algorithm is not in operation for a very long time. The nature of the 'convexification' should often reduce the magnitude of this problem, and 'robustify' the process. In a sense, the decentralized algorithm would perform much better than the worst of q-identical (but not communicating) processors, and (in a tracking system, for example) would reduce the chances of any one processor losing track. In applications to optimization or systems evaluation by monte-carlo simulation one can use 'variance reduction' ideas in choosing appropriate correlations among the sets $\{\xi_n^i, n = 1,2, ...\}$, $i \leqslant q$. Hopefully this, together with the above 'robustifying' property, would yield good behavior.

An Example. The following is an example which opens up many new possibilities. Consider two receivers -- say, digital phase locked loops -- each receiving a signal from the same source, but the two being physically separated. Each must estimate the phase or epoch of the signal pulse (and perhaps the phase of the carrier). Suppose that the source is much farther from the receivers than they are to each other, so that more reliable communication between the receivers is possible. It might be possible to improve each others estimates by occasional communcations. This communication would transfer the estimates -- as well as allow the receivers to improve the mutual synchronization of their clocks or oscillators -- so that the transferred estimates can be meaningfully used.

Communication Noise. In examples such as the preceeding, one would normally have communication noise. This is readily incorporated into the analysis. Write (1.1) as

$$(7.1) \qquad X_{n+1} = A_n(X_n + \check{\delta}_n) + \epsilon B(X_n,\xi_n),$$

where $\check{\delta}_n$ represents the communication noise. For the algorithm to be useful at all, this noise should be of an order no larger than ϵ. Then write $\check{\delta}_n = \epsilon\delta_n$, and proced as before.

Even if $\check{\delta}_n = 0(\sqrt{\epsilon})$ and $E\check{\delta}_n = 0$, useful results can be obtained. If the interpolation of

$$\left\{ \sum_{k=0}^{n} \Phi(n|k)\check{\delta}_k \right\}$$

converges weakly to a Wiener process $\check{W}(\cdot)$, then we might have $X^{\epsilon}(\cdot) \Rightarrow X(\cdot)$:

$$dX = \bar{\Phi}\,\bar{B}(X)dt + d\check{W}.$$

Again $X(\cdot)$ takes the form $(x(\cdot), \ldots, x(\cdot))$, under appropriate conditions on $\{\check{\delta}_n\}$.

An Alternative Algorithm. To get additional insight into the behavior of decentralized algorithm, we formally compare (1.1) with a reasonable alternative. Suppose that the processors communicate and 'convexify' only the changes in the states since the last communication. In particular, let $q = 2$ and let $\{\tau^i_n\}$, $i = 1,2$, denote the comunication times of the two processors, with $|\tau^i_{n+1} - \tau^i_n|$ bounded. Here processor 2 communicates to processor 1 at $\{\tau^1_n\}$, and similarly for the reverse communication.. We proceed purely formally, and suppose that the dynamics are smooth and bounded. Define $\{\tilde{X}^i_n\}$ by $\tilde{X}^i_{\tau^i_k} = X^i_{\tau^i_k}$ and

$$(7.2) \qquad \check{X}^i_{n+1} = \check{X}^i_n + \epsilon b^i(\check{X}^i_n, \xi^i_n), \qquad \tau^i_k \leqslant n < \tau^i_{k+1}.$$

For $\alpha \epsilon (0,1/2]$, set

$$(7.3) \qquad \begin{aligned} X^1_{\tau^1_{k+1}} &= X^1_{\tau^1_k} + (1-\alpha)\epsilon \sum_{\tau^1_k}^{\tau^1_{k+1}-1} b^1(\check{X}^1_n, \xi^1_n) + \alpha\epsilon \sum_{\tau^1_k}^{\tau^1_{k+1}-1} b^2(\check{X}^2_n, \xi^2_n) \\ X^2_{\tau^2_{k+1}} &= X^2_{\tau^2_k} + \alpha\epsilon \sum_{\tau^2_k}^{\tau^2_{k+1}-1} b^1(\check{X}^1_n, \xi^1_n) + (1-\alpha)\epsilon \sum_{\tau^2_k}^{\tau^2_{k+1}-1} b^2(\check{X}^2_n, \xi^2_n). \end{aligned}$$

Owing to the smoothness and boundedness assumptions, there are $0^i_n(\epsilon^2) = 0(\epsilon^2)$ and a process $\hat{X}_n = (\hat{X}^1_n, \hat{X}^2_n)$ satisfying (7.4) and which equals (modulo $0(\epsilon^2)$) $(\check{X}^1_n, \check{X}^2_n)$ and $(X^1_{\tau^1_k}, X^2_{\tau^2_k})$ (at the communication times)

$$(7.4)\qquad \begin{aligned} \hat{X}^1_{n+1} &= \hat{X}^1_n + (1-\alpha)\epsilon b^1(\hat{X}^1_n,\xi^1_n) + \alpha\epsilon b^2(\hat{X}^2_n,\xi^2_n) + 0^1_n(\epsilon^2) \\ \hat{X}^2_{n+1} &= \hat{X}^2_n + \alpha\epsilon b^1(\hat{X}^1_n,\xi^1_n) + (1-\alpha)\epsilon b^2(\hat{X}^2_n,\xi^2_n) + 0^2_n(\epsilon^2). \end{aligned}$$

The 'size' of the 0^i_n depend on the bound on $|\tau^i_{k+1} - \tau^i_k|$. From this point on, one can use standard theory for the centralized case to get both the ODE and the asymptotic normalized variance. Define $\hat{X}^\epsilon(\cdot)$ as $X^\epsilon(\cdot)$ was defined, and similarly for $\hat{U}^\epsilon(\cdot)$. The limit ODE is

$$(7.5)\qquad \dot{\hat{X}} = \begin{matrix} (1-\alpha)\bar{b}^1(\hat{X}^1) + \alpha\bar{b}^2(\hat{X}^2) \\ \alpha\bar{b}^1(\hat{X}^1) + (1-\alpha)\bar{b}^2(\hat{X}^2) \end{matrix} = \hat{B}(\hat{X}^1,\hat{X}^2)$$

The limit $\hat{U}(\cdot)$ of $\{\hat{U}^\epsilon(\cdot)\}$ satisfies

$$(7.6)\qquad d\hat{U} = \hat{M}\hat{U}\,dt + d\hat{W}\ ,$$

where

$$\text{cov }\hat{W}(1) = \sum_{-\infty}^{\infty} E\bar{\xi}_n\bar{\xi}_0'\ ,$$

$$\bar{\xi}_n = \begin{bmatrix} (1-\alpha)b^1(\theta,\xi^1_n) + \alpha b^2(\theta,\xi^2_n) \\ \alpha b^1(\theta,\xi^1_n) + (1-\alpha)b^2(\theta,\xi^2_n) \end{bmatrix},$$

$$\hat{M} = (\hat{B}(\theta,\theta))_X\ ,$$

and we suppose that $\hat{M}$ is a stable matrix.

Comparison of the Alternative (7.2), (7.3) with (1.1). We use the special scalar case of the first subsection of this section, where $\{\xi^i_n\}$ are i.i.d. and $b^1(\cdot,\cdot) = b^2(\cdot,\cdot) = b(\cdot,\cdot)$. Then (again $\bar{b}_x(\theta) < 0$)

$$\dot{\hat{X}} = \begin{Bmatrix} (1-\alpha)\bar{b}(\hat{X}^1) + \alpha\bar{b}(\hat{X}^2) \\ \\ \alpha\bar{b}(\hat{X}^1) + (1-\alpha)\bar{b}(\hat{X}^2) \end{Bmatrix}$$

$$d\hat{U} = \bar{b}_x(\theta) \begin{bmatrix} (1-\alpha) & \alpha \\ \alpha & (1-\alpha) \end{bmatrix} \hat{U}\, dt + d\hat{W} = M\hat{U}dt + d\hat{W}$$

$$\text{cov } \hat{W}(1) = Eb^2(\theta,\xi_n^i) \begin{bmatrix} (1-\alpha)^2 + \alpha^2 & 2\alpha(1-\alpha) \\ 2\alpha(1-\alpha) & (1-\alpha)^2 + \alpha^2 \end{bmatrix}$$

Let var_{D2} denote the stationary variance of $\hat{U}(\cdot)$. As $\alpha \uparrow 1/2$, this converges to the infimal value, equal to var_{2C}. But at $\alpha = 1/2$, the matrix M is singular. Thus, again, there seems to be a trade-off between the 'minimal asymptotic variance' and the length of time one must wait for the asymptotic estimates to be valid or, similarly, for the communication to be effective. At this point, the alternative algorithm does not seem to offer any clear advantages. It was investigated simply because of the idea that there might be an advantage in communicating only recent data.

8. Stochastic Approximation With $\epsilon_n \to 0$

The entire development can be repeated if ϵ is replaced by $0 < \epsilon_n \to 0$, $\Sigma\epsilon_n = \infty$. One then gets results of classical stochastic approximation type, and we only make a few formal comments. We use $X_{n+1} = A_nX_n + \epsilon_n b(X_n,\xi_n)$. Define $t_n = \Sigma_0^{n-1}\epsilon_i$ and (for $t \geq 0$) define $X^n(\cdot)$ by $X^n(t) = X_{n+i}$ for $t \in [t_i-t_n, t_{i+1}-t_n)$. Under the conditions of Theorem 5.1, $\overline{\lim}_n EV(X_n) < \infty$. Given either this or the use of the projection algorithm of Section 4, one can get the appropriate ODE which characterizes the limit paths. If this has the appropriate stability properties (as in Section 5), we can show that $X^{\epsilon,i}(\cdot) \Rightarrow x(\cdot) \equiv \theta$. The ODE is the same as that in the previous sections, for all the same cases.

If $\Sigma\epsilon_n^2 < \infty$, then the idea in [10] can be adapted to get w.p.1 convergence results.

Appendix. Some Results in Weak Convergence.

For some integer s, let $D[0,\infty)$ denote the space of E^s-valued functions on $[0,\infty)$ which are right continuous and have left hand limits, with the Skorohod topology [7, Chapter 2]. This topology is defined as follows. Let A be the set of strictly increasing Lipschitz continuous functions from $[0,\infty)$ onto $[0,\infty)$. Define the metric

$$d(x(\cdot),y(\cdot)) = \inf_{\lambda \in A} \max\left\{ \sup_{s>t\geq 0} \left| \log \left[\frac{\lambda(s)-\lambda(t)}{s-t} \right] \right|, \int_0^\infty e^{-\tau} d_\tau(x(\cdot),y(\cdot),\lambda) d\tau \right\},$$

where $d_\tau(x(\cdot),y(\cdot),\lambda) = \min(1, \sup_t |x(\lambda(t) \cap \tau) - y(\lambda(t) \cap \tau)|)$.

Define $\{Z_n^\epsilon\}$ and $\{Z^\epsilon(\cdot)\}$ by $Z_{n+1}^\epsilon = Z_n^\epsilon + \epsilon F_n^\epsilon$, $Z^\epsilon(t) = Z_n^\epsilon$ $[n\epsilon, n\epsilon+\epsilon)$. If $\{Z_0^\epsilon\}$ is tight and the $\{F_n^\epsilon\}$ are uniformly integrable, then $\{Z^\epsilon(\cdot)\}$ is tight in $D[0,\infty)$ and all weak limits are absolutely continuous.

Let $Z^\epsilon(\cdot) \Rightarrow Z(\cdot)$ in $D[0,\infty)$. By a suitable choice of the probability space, the weak convergence becomes convergence w.p.1 in the metric of $D[0,\infty)$ [13, Theorem 3.1.1]. I.e., there is a probability space $(\tilde{\Omega},\tilde{B},\tilde{P})$ with processes $\{\tilde{Z}^\epsilon(\cdot)\}$, $\tilde{Z}(\cdot)$ defined on it so that for each Borel set B in $D[0,\infty)$, $\tilde{P}\{\tilde{Z}^\epsilon(\cdot) \in B\} = P\{Z^\epsilon(\cdot) \in B\}$, $\tilde{P}\{\tilde{Z}(\cdot) \in B\} = P\{Z(\cdot) \in B\}$ and $\tilde{Z}^\epsilon(\cdot) \to \tilde{Z}(\cdot)$ w.p.1 in the topology of $D[0,\infty)$. The use of this representation often facilitates the analysis and characterization of the limits.

References

[1] J.N. Tsitsiklis, "Problems in Decentralized Decision Making and Computation," Ph.D. Thesis, Electrical Engineering Deptartment, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[2] D. Bertsekas, J.N. Tsitsiklis, M. Athens, "Convergence Theories of Distributed Iterative Processes: A Survey," Technical report for Information nd Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[3] A.P. Korostelev, Stochastic Recurrent Processes, 1984, Nauka, Moscow.

[4] P. Dupuis and H.J. Kushner, "Stochastic Approximations via Large Deviations: Asymptotic Properties," SIAM J. on Control and Optimization, September 1985.

[5] H.J. Kushner and Hai Huang, "Averaging Methods for the Asymptotic Analysis of Learning and Adaptive Systems with Small Adjustment Rate," SIAM J. on Control and Optimization, 19 (1981), 635-650.

[6] P. Billingsley, Convergence of Probability Measures, 1968, Wiley, New York.

[7] T.G. Kurtz, Approximation of Population Processes, 1981, Vol. 36 in CBMS-NSF Regional Conf. Series in Appl. Math., Soc. for Ind. and Appl. Math.

[8] H.J. Kushner, Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory, M.I.T. Press, Cambridge, MA, 1984.

[9] H.J. Kushner and A. Shwartz, "An Invariant Measure Approach to the Convergence of Stochastic Approximations with State-dependent Noise," SIAM J. on Control and Optimization, 22, January 1984, 13-27.

[10] H.J. Kushner, "An Averaging Method for Stochastic Approximations with Discontinuous Dynamics, Constraints, and State-dependent Noise," in Recent Advances in Stochastics, Ed. Rizri, Rustagi and Siegmund, Academic Press (1983).

[11] G. Blankenship and G.C. Papanicolaou, "Stability and Control of Stochastic Systems with Wide Band Noise Disturbances," SIAM J. Appl. Math., 34 (1978), 437-476.

[12] H.J. Kushner and Hai Huang, "Asymptotic Properties of Stochastic Approximations with Constant Coefficients," SIAM J. on Control and Optimization, 19 (1981), 87-105.

[13] A.V. Skorohod, "Limit Theorems for Stochastic Processes," Theory of Probability and its Applications, 1, (1956), 262-290.

END

1-87

DTIC